EP04/11974

Europäisches Patentamt
European Patent Office
Office européen des brevets

REC'D 11 JAN 2005
WIPO PCT

Bescheinigung Certificate Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. Patent application No. Demande de brevet n°

04002722.9

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

Der Präsident des Europäischen Patentamts
Im Auftrag

For the President of the European Patent Of

Le Président de l'Office européen des breve
p.o.

R C van Dijk

BEST AVAILABLE COPY

Anmeldung Nr:
Application no.: 04002722.9
Demande no:

Anmeldetag:
Date of filing: 06.02.04
Date de dépôt:

Anmelder/Applicant(s)/Demandeur(s):

Selexis S.A.
18, chemin des Aulx
1228 Plan-les-Quates/Geneva
SUISSE

Bezeichnung der Erfindung/Title of the invention/Titre de l'invention:
(Falls die Bezeichnung der Erfindung nicht angegeben ist, siehe Beschreibung.
If no title is shown please refer to the description.
Si aucun titre n'est indiqué se referer à la description.)

High efficiency gene transfer and expression in mammalian cells by a multiple transfection procedure of MAR sequences

In Anspruch genommene Priorität(en) / Priority(ies) claimed /Priorité(s) revendiquée(s)
Staat/Tag/Aktenzeichen/State/Date/File no./Pays/Date/Numéro de dépôt:

Internationale Patentklassifikation/International Patent Classification/
Classification internationale des brevets:

C12N15/00

Am Anmeldetag benannte Vertragstaaten/Contracting states designated at date of filing/Etats contractants désignées lors du dépôt:

AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HU IE IT LU MC NL PT RO SE SI SK TR LI

HIGH EFFICIENCY GENE TRANSFER AND EXPRESSION IN MAMMALIAN CELLS BY A MULTIPLE TRANSFECTION PROCEDURE OF MAR SEQUENCES

FIELD OF THE INVENTION

The present invention relates to purified and isolated DNA sequences having protein production increasing activity and more specifically to the use of matrix attachment regions (MARs) for increasing protein production activity in a eukaryotic cell. Also disclosed is a method for the identification of said active regions, in particular MAR nucleotide sequences, and the use of these characterized active MAR sequences in a new multiple transfection method.

BACKGROUND OF THE INVENTION

Nowadays, the model of loop domain organization of eukaryotic chromosomes is well accepted (Boulikas T, "Nature of DNA sequences at the attachment regions of genes to the nuclear matrix", *J. Cell Biochem.*, 52:14-22, 1993). According to this model chromatin is organized in loops that span 50-100 kb attached to the nuclear matrix, a proteinaceous network made up of RNPs and other nonhistone proteins (Bode J, Stengert-Iber M, Kay V, Schalke T and Dietz-Pfeilstetter A, *Crit. Rev. Euk. Gene Exp.*, 6:115-138, 1996).

The DNA regions attached to the nuclear matrix are termed SAR or MAR for respectively scaffold (during metaphase) or matrix (interphase) attachment regions (Hart, C., and Laemmli, U. (1998), "Facilitation of chromatin dynamics by SARs" *Curr Opin Genet Dev* 8, 519-525.)

As such, these regions may define boundaries of independent chromatin domains, such that only the encompassing cis-regulatory elements control the expression of the genes within the domain.

However, their ability to fully shield a chromosomal locus from nearby chromatin elements, and thus confer position-independent gene expression, has not been seen in stably transfected cells (Poljak, L., Seum, C.,Mattioni, T., and Laemmli, U. (1994) "SARs stimulate but do not confer position independent gene expression", *Nucleic*

Acids Res 22, 4386-4394). On the other hand, MAR sequences have been shown to interact with enhancers to increase local chromatin accessibility (Jenuwein, T., Forrester, W., Fernandez-Herrero, L., Laible, G., Dull, M., and Grosschedl, R. (1997) "Extension of chromatin accessibility by nuclear matrix attachment regions" *Nature* 385, 269-272). Specifically, MAR elements can enhance expression of heterologous genes in cell culture lines (Kalos, M., and Fournier, R. (1995) "Position-independent transgene expression mediated by boundary elements from the apolipoprotein B chromatin domain" *Mol Cell Biol* 15,198-207), transgenic mice (Castilla, J., Pintado, B., Sola, I., Sanchez-Morgado, J., and Enjuanes, L. (1998) "Engineering passive immunity in transgenic mice secreting virus-neutralizing antibodies in milk" *Nat Biotechnol* 16, 349-354) and plants (Allen, G., Hall, G. J., Michalowski, S., Newman, W., Spiker, S., Weissinger, A., and Thompson, W. (1996), "High-level transgene expression in plant cells: effects of a strong scaffold attachment region from tobacco" *Plant Cell* 8, 899-913). The utility of MAR sequences for developing improved vectors for gene therapy is also recognized (Agarwal, M., Austin, T., Morel, F., Chen, J., Bohnlein, E., and Plavec,I. (1998), "Scaffold attachment region-mediated enhancement of retroviral vector expression in primary T cells" *J Virol* 72, 3720-3728).

Recently, it has been shown that the chicken lysozyme 5' MAR is able to significantly enhance reporter expression in pools of stable Chinese Hamster Ovary (CHO) cells (Zahn-Zabal, M., et al., "Development of stable cell lines for production or regulated expression using matrix attachment regions" *J Biotechnol*, 2001, 87(1): p. 29-42). This property was used to increase the proportion of high-producing clones, thus reducing the number of clones that need to be screened. These benefits have been observed both for constructs with MARs flanking the transgene expression cassette, as well as when constructs are co-transfected with the MAR on a separate plasmid. However, expression levels upon co-transfection with MARs were not as high as those observed for a construct in which two MARs delimit the transgene expression unit. Another limitation of this technique is the quantity of DNA that can be transfected per cell. Many multiples transfection protocols have been developed in order to achieve a high transfection efficiency to characterize the function of genes of interest. The protocol applied by Yamamoto et al, 1999 ("High efficiency gene transfer by multiple transfection protocol", *Histochem. J.* 31(4), 241-243) leads to a transfection efficiency of about 80 % after 5 transfections events, whereas the conventional transfection protocol only achieved a rate of <40%. While this technique may be useful when one wishes to

increase the proportion of expressing cells, it does not lead to cells with a higher intrinsic productivity. Therefore, it cannot be used to generate high producer monoclonal cell lines. Hence, the previously described technique has two major drawbacks:

i) this technique does not generate a homogenous population of transfected cells, since it cannot favour the integration of further gene copy, nor does it direct the transgenes to favorable chromosomal loci,

ii) the use of the same selectable marker in multiple transfection events does not permit the selection of doubly or triply transfected cells.

In patent application WO02/074969, the utility of MARs for the development of stable eukaryotic cell lines has also been demonstrated. However, this application does not disclose neither any conserved homology for MAR DNA element nor any technique for predicting the ability for a DNA sequence to be a MAR sequence.

In fact no clear-cut MAR consensus sequence has been found (Boulikas T, "Nature of DNA sequences at the attachment regions of genes to the nuclear matrix", *J. Cell Biochem.*, 52:14-22, 1993) but evolutionarily, the structure of these sequences seem to be functionally conserved in eukaryotic genomes, since animal MARs can bind to plant nuclear scaffolds and vice versa (Mielke C, Kohwi Y, Kohwi-Shigematsu T and Bode J, "Hierarchical binding of DNA fragments derived from scaffold-attached regions: correlation of properties in vitro and function in vivo", *Biochemistry*, 29:7475-7485, 1990) .

The identification of MARs by biochemical studies is a long and unpredictable process; various results can be obtained depending on the assay (Razin SV, "Functional architecture of chromosomal DNA domains", *Crit Rev Eukaryot Gene Expr.*, 6:247-269, 1996). Considering the huge number of expected MARs in a eukaryotic genome and the amount of sequences issued from genome projects, a tool able to filter potential MARS in order to perform targeted experiments would be greatly useful.

Currently two different predictive tools for MARs are available via the Internet. The fist one, MAR-Finder (http://futuresoft.org/MarFinder; Singh GB, Kramer JA and Krawetz SA, "Mathematical model to predict regions of chromatin attachment to the nuclear matrix", *Nucleic Acid Research*, 25:1419-1425, 1997) is based on set of

patterns identified within several MARs and a statistical analysis of the co-occurrence of these patterns. MAR-Finder predictions are dependent of the sequence context, meaning that predicted MARs depend on the context of the submitted sequence. The other predictive software, SMARTest (http://www. genomatix.de; Frisch M, Frech K, Klingenhoff A, Cartharius K, Liebich I and Werner T, "In silico prediction of scaffold/matrix attachment regions in large genomic sequences", *Genome Research*, 12:349-354, 2001), use weight-matrices derived from experimentally identified MARs. SMARTest is said to be suitable to perform large-scale analyses. But actually aside its relative poor specificity, the amount of hypothetical MARs rapidly gets huge when doing large scale analyses with it, and in having no way to increase its specificity to restrain the number of hypothetical MARs, SMARTest becomes almost useless to plan efficiently wet-lab experiments.

Some other softwares, not available via the Internet, also exists; they are based as well on the frequency of MAR motifs (MRS criterion;Van Drunen CM et al., "A bipartite sequence element associated with matrix/scaffold attachment regions", *Nucleic Acids Res*, 27:2924-2930, 1999), (ChrClass; Glazko GV et al., "Comparative study and prediction of DNA fragments associated with various elements of the nuclear matrix", *Biochim. Biophys. Acta*, 1517:351-356, 2001) or based on the identification of sites of stress-induced DNA duplex (SIDD; Benham C and al., "Stress-induced duplex DNA destabilization in scaffold/matrix attachment regions", *J. Mol. Biol.*, 274:181-196, 1997). However, their suitability to analyze complete genome sequences remains unknown, and whether these tools may allow the identification of protein production-increasing sequences has not been reported.

All the above available predictive methods have some drawbacks that prevent large-scale analyses of genomes to identify reliably novel and potent MARs.

SUMMARY OF THE INVENTION

Therefore, the object of the present invention is to provide an improved method for the identification of DNA sequences having protein production increasing activity, in particular MAR nucleotide sequences, and the use of these characterized active MAR sequences in a new multiple transfection method to increase the production of recombinant proteins in eukaryotic cells.

BRIEF DESCRIPTION OF THE FIGURES

Fig. 1 shows the distribution plots of MARs and non-MARs sequences. Histograms are density plots (relative frequency divided by the bin width) relative to the score of the observed parameter. The density histogram for human MARs in the SMARt DB database is shown in black, while the density histogram for the human chromosome 22 are in grey.

Fig. 2 shows Scatterplots of the four different criteria used by SMAR Scan and the AT-content with human MARs from SMARt DB.

Fig. 3 shows the distribution plots of MAR sequences by organism. MAR sequences from SMARt DB of other organisms were retrieved and analyzed. The MAR sequences density distributions for the mouse, the chicken, the sorghum bicolor and the human are plotted jointly.

Fig. 4 shows SMAR Scan predictions on human chromosome 22 and onshuffled chromosome 22. Left plot corresponds to the number of hits obtained by SMAR Scan when analyzing crumbled, scrambled and native chromosome 22. Right plot represents the number of S/MARs predicted by SMAR Scan in crumbled, scrambled and native chromosome 22.

Fig. 5 shows the dissection of the ability of the chicken lysozyme gene 5'-MAR to stimulate transgene expression in CHO-DG44 cells. Fragments B, K and F show the highest ability to stimulate transgene expression. The indicated relative strength of the elements was based on the number of high-expressor cells.

Fig. 6 shows the effect of serial-deletions of the 5'-end (upper part) and the 3'-end (lower part) of the 5'-MAR on the loss of ability to stimulate transgene expression. The transition from increased to decreased activity coincide with B-, K- and F-fragments.

Fig. 7 shows that portions of the F fragment significantly stimulate transgene expression. The F fragment regions indicated by the light grey arrow were multimerized, inserted in pGEGFP Control and transfected in CHO cells. The element that displays the highest activity is located in the central part of the element and corresponds to

fragment FIII (black bar labelled minimal MAR). In addition, an enhancer activity is located in the 3'-flanking part of the FIII fragment (dark grey bar labelled MAR enhancer).

Fig. 8 shows a map of locations for various DNA sequence motifs within the *cLysMAR*. Fig. 8 (B) represents a Map of locations for various DNA sequence motifs within the cLysMAR. Vertical lines represent the position of the computer-predicted sites or sequence motifs along the 3034 base pairs of the cLysMAR and its active regions, as presented in Fig. 5. The putative transcription factor sites, (MEF2 05, Oct-1, USF-02, GATA, NFAT) for activators and (CDP, SATB1, CTCF, ARBP/MeCP2) for repressors of transcription, were identified using MatInspector (Genomatix), and CpG islands were identifed with CPGPLOT. Motifs previously associated with MAR elements are labelled in black and include CpG dinucleotides and CpG islands, unwinding motifs (AATATATT and AATATT), poly As and Ts, poly Gs and Cs, Drosophila topoisomerase II binding sites (GTNWAYATTNATTNATNNR) which had identity to the 6 bp core and High mobility group I (HMG-I/Y) protein binding sites. Other structural motifs labelled in include nucleosome-binding and nucleosome disfavouring sites and a motif thought to relieve the superhelical strand of DNA. Fig. 8(A) represents the comparison of the ability of portions of the cLysMAR to activate transcription with MAR prediction score profiles with MarFinder. The top diagram shows the MAR fragment activity as in Fig. 5, while the middle and bottom curves show MARFinder-predicted potential for MAR activity and for bent DNA structures respectively.

Fig. 9 shows the Correlation of DNA physico-chemical properties with MAR activity. Fig. 9(A), represents the DNA melting temperature, double helix bending, major groove depth and minor groove width profiles of the 5'-MAR and were determined using the algorithms of Levitsky et al. The most active B, K and F fragments depicted at the top are as shown as in Figure 1. Fig. 9(B), represents the enlargement of the data presented in panel A to display the F fragment map aligned with the tracings corresponding to the melting temperature (top curve) and DNA bending (bottom curve). The position of the most active FIB fragment and protein binding site for specific transcription factors are as indicated.

Fig. 10 shows the distribution of putative transcription factor binding sites within the 5'-MAR. Large arrows indicate the position of the CUE elements as identified with SMARscan.

Fig. 11 shows the scheme of assembly of various portions of the MAR. The indicated portions of the cLysMAR were amplified by PCR, introducing BglII-BamHI linker elements at each extremity, and assembled to generate the depicted composite elements. For instance, the top construct consists of the assembly of all CUE and flanking sequences at their original location except that BglI-BamHII linker sequences separate each element.

Fig. 12 represents the plasmid maps.

Fig. 13 shows the effect of re-transfecting primary transfectants on GFP expression. Cells (CHO-DG44) were co-transfected with pSV40EGFP (left tube) or pMAR-SV40EGFP (central tube) and pSVneo as resistance plasmid. Cells transfected with pMAR-SV40EGFP were re-transfected 24 hours later with the same plasmid and a different selection plasmid, pSVpuro (right tube). After two weeks selection, the phenotype of the stably transfected cell population was analysed by FACS.

Fig. 14 shows the effect of multiple load of MAR-containing plasmid. The pMAR-SV40EGFP/ pMAR-SV40EGFP secondary transfectants were used in a third cycle of transfection at the end of the selection process. The tertiary transfection was accomplished with pMAR or pMAR-SV40EGFP to give tertiary transfectants. After 24 hours, cells were transfected again with either plasmid, resulting in the quaternary transfectants (see Table 4).

DETAILED DESCRIPTION OF THE INVENTION

The present invention relates to a purified and isolated DNA sequence having protein production increasing activity characterized in that said DNA sequence comprises at least one bent DNA element, and at least one binding site for a DNA binding protein.

Certain sequences of DNA are known to form a relatively "static curve", where the DNA follows a particular 3-dimensional path. Thus, instead of just being in the normal B-

DNA conformation ("straight"), the piece of DNA can form a flat, planar curve also defined as bent DNA (Marini, *et al.*, 1982 "Bent helical structure in kinetoplast DNA", *Proc. Natl. Acad. Sci. USA*, 79: 7664-7664).

According to the present invention, the bent DNA element is usually a MAR nucleotide sequence selected from the group comprising the sequences SEQ ID Nos 1 to 23 or a cLysMAR element or a fragment thereof. Preferably, the bent DNA element is a MAR nucleotide sequence selected from the group comprising the sequences SEQ ID Nos 1 to 23, more preferably the sequences SEQ ID Nos 21 to 23.

Encompassed by the present invention are as well complementary sequences of the above-mentioned sequences SEQ ID Nos 1 to 23 and the cLysMAR element or fragment, which can be produced by using PCR

An element is a conserved nucleotide sequences that bears common functional properties (i.e. binding sites for transcription factors) or structural (i.e. bent DNA sequence) features.

A part of sequences SEQ ID Nos 1 to 23 and the cLysMAR element or fragment refers to sequences sharing at least 70% nucleotides in length with the respective sequence of the SEQ ID Nos 1 to 23. These sequences can be used as long as they exhibit the same properties as the native sequence from which they derive. Preferably these sequences share more than 80%, in particular more than 90% nucleotides in length with the respective sequence of .the SEQ ID Nos 1 to 23.

The present invention also includes variants of the aforementioned sequences SEQ ID Nos 1 to 23 and the cLysMAR element or fragment, that is nucleotide sequences that vary from the reference sequence by conservative nucleotide substitutions, whereby one or more nucleotides are substituted by another with same characteristics.

The sequences SEQ ID Nos 1 to 23 have been identified by scanning human chromosome 1 and 2 using SMAR Scan, showing that the identification of novel MAR sequences is feasible using the tools reported thereafter.

In a first step, the complete chromosome 1 and 2 were screened to identify bent DNA element as region corresponding to the highest bent, major groove depth, minor groove width and lowest melting temperature as in figure 3. In a second step, this collection of sequence was scanned for binding sites of regulatory proteins such as SATB1, GATA, etc. as shown in the figure 8B) yielding sequences SEQ ID 1-23. Furthermore, sequences 21-23 were further shown to be located next to known gene from the Human Genome Data Base.

Molecular chimera of MAR sequences are also considered in the present invention. By molecular chimera is intended a nucleotide sequence that may include a functional portion of a MAR element and that will be obtained by molecular biology methods known by those skilled in the art.

Particular combinations of MAR elements or fragments or sub-portions thereof are also considered in the present invention. These fragments can be prepared by a variety of methods known in the art. These methods include, but are not limited to, digestion with restriction enzymes and recovery of the fragments, chemical synthesis or polymerase chain reactions (PCR).
Therefore, particular combinations of elements or fragments of the sequences SEQ ID Nos 1 to 23 and cLysMAR elements or fragments are also envisioned in the present invention, depending on the functional results to be obtained. Elements of the cLysMAR are e.g. the B, K and F regions as described in WO 02/074969, the disclosure of which is hereby incorporated herein by reference, in its entirety. The preferred elements of the cLysMAR used in the present invention are the B, K and F regions. Only one element might be used or multiple copies of the same or distinct elements (multimerized elements) might be used (see Fig. 8 A)).

By fragment is intended a portion of the respective nucleotide sequence. Fragments of a MAR nucleotide sequence may retain biological activity and hence bind to purified nuclear matrices and/or alter the expression patterns of coding sequences operably linked to a promoter. Fragments of a MAR nucleotide sequence may range from at least about 100 to 1000 bp, preferably from about 200 to 700 bp, more preferably from about 300 to 500 bp nucleotides. Also envisioned are any combinations of fragments, which have the same number of nucleotides present in a synthetic MAR sequence

consisting of natural MAR element and/or fragments. The fragments are preferably assembled by linker sequences. Preferred linkers are BglI-BamHI linker.

"Protein production increasing activity" refers to an activity of the purified and isolated DNA sequence defined as follows: after having been introduced under suitable conditions into a eukaryotic host cell, the sequence is capable of increasing protein production levels in cell culture as compared to a culture of cell transfected without said DNA sequence. Usually the increase is 1.5 to 6 fold, preferably 4 to 6 fold. This corresponds to a production rate or a specific cellular productivity of at least 10 pg per cell per day (see Example 11 and Fig.13).

As used herein, the following definitions are supplied in order to facilitate the understanding of this case. To the extent that the definitions vary from meanings circulating within the art, the definitions below are to control.

"Chromatin" is the nucleic acid material having the chromosomes of a eukaryotic cell, and refers to DNA, RNA and associated proteins.

A "chromatin element" means a nucleic acid sequence on a chromosome.

"Cis" refers to the placement of two or more elements (such as chromatin elements) on the same nucleic acid molecule (such as the same vector or chromosome).

"Trans" refers to the placement of two or more elements (such as chromatin elements) on two or more different nucleic acid molecules (such as on two vectors or two chromosomes).

Chromatin modifying elements that are potentially capable of overcoming position effects, and hence are of interest for the development of stable cell lines, include boundary elements (BEs), matrix attachment regions (MARs), locus control regions (LCRs), and universal chromatin opening elements (UCOEs).

Boundary elements ("BEs"), or insulator elements, define boundaries in chromatin in many cases (Bell, A., and Felsenfeld, G. 1999; "Stopped at the border: boundaries and insulators, *Curr Opin Genet Dev* 9, 191-198) and may play a role in defining a

transcriptional domain in vivo. BEs lack intrinsic promoter/enhancer activity, but rather are thought to protect genes from the transcriptional influence of regulatory elements in the surrounding chromatin. The enhancer-block assay is commonly used to identify insulator elements. In this assay, the chromatin element is placed between an enhancer and a promoter, and enhancer-activated transcription is measured. Boundary elements have been shown to be able to protect stably transfected reporter genes against position effects in Drosophila, yeast and in mammalian cells. They have also been shown to increase the proportion of transgenic mice with inducible transgene expression.

Locus control regions ("LCRs") are cis-regulatory elements required for the initial chromatin activation of a locus and subsequent gene transcription in their native locations (Grosveld, F. 1999, "Activation by locus control regions?" *Curr Opin Genet Dev* 9, 152-157). The activating function of LCRs also allows the expression of a coupled transgene in the appropriate tissue in transgenic mice, irrespective of the site of integration in the host genome. While LCRs generally confer tissue-specific levels of expression on linked genes, efficient expression in nearly all tissues in transgenic mice has been reported for a truncated human T-cell receptor LCR and a rat LAP LCR. The most extensively characterized LCR is that of the globin locus. Its use in vectors for the gene therapy of sickle cell disease and (3-thalassemias is currently being evaluated.

Ubiquitous chromatin opening elements ("UCOEs", also known as "ubiquitously-acting chromatin opening elements") have been reported in WO 00/05393.

An "enhancer" is a nucleotide sequence that acts to potentiate the transcription of genes independent of the identity of the gene, the position of the sequence in relation to the gene, or the orientation of the sequence. The vectors of the present invention optionally include enhancers.

A "gene" is a deoxyribonucleotide (DNA) sequence coding for a given mature protein. As used herein, the term "gene" shall not include untranslated flanking regions such as RNA transcription initiation signals, polyadenylation addition sites, promoters or enhancers.

A "product gene" is a gene that encodes a protein product having desirable

characteristics such as diagnostic or therapeutic utility. A product gene includes, e. g., structural genes and regulatory genes.

A "structural gene" refers to a gene that encodes a structural protein. Examples of structural genes include but are not limited to, cytoskeletal proteins, extracellular matrix proteins, enzymes, nuclear pore proteins and nuclear scaffold proteins, ion channels and transporters, contractile proteins, and chaperones. Preferred structural genes encode for antibodies or antibody fragments.

A "regulatory gene" refers to a gene that encodes a regulatory protein. Examples of regulatory proteins include, but are not limited to, transcription factors, hormones, growth factors, cytokines, signal transduction molecules, oncogenes, proto-oncogenes, transmembrane receptors, and protein kinases.

"Orientation" refers to the order of nucleotides in a given DNA sequence. For example, an inverted orientation of a DNA sequence is one in which the 5' to 3' order of the sequence in relation to another sequence is reversed when compared to a point of reference in the DNA from which the sequence was obtained. Such reference points can include the direction of transcription of other specified DNA sequences in the source DNA and/or the origin of replication of replicable vectors containing the sequence.

"Eukaryotic cell" refers to any mammalian or non-mammalian cell from a eukaryotic organism. By way of non-limiting example, any eukaryotic cell that is capable of being maintained under cell culture conditions and subsequently transfected would be included in this invention. Especially preferable cell types include, e. g., stem cells, embryonic stem cells, Chinese hamster ovary cells (CHO), COS, BHK21, NIH3T3, HeLa, C2C12, cancer cells, and primary differentiated or undifferentiated cells. Other suitable host cells are known to those skilled in the art.

The terms "host cell" and "recombinant host cell" are used interchangeably herein to indicate a eukaryotic cell into which one or more vectors of the invention have been introduced. It is understood that such terms refer not only to the particular subject cell but also to the progeny or potential progeny of such a cell. Because certain modifications may occur in succeeding generations due to either mutation or

environmental influences, such progeny may not, in fact, be identical to the parent cell, but are still included within the scope of the term as used herein.

The terms "introducing a purified DNA into a eukaryotic host cell" or "transfection" denote any process wherein an extracellular DNA, with or without accompanying material, enters a host cell. The term "cell transfected" or "transfected cell" means the cell into which the extracellular DNA has been introduced and thus harbours the extracellular DNA. The DNA might be introduced into the cell so that the nucleic acid is replicable either as a chromosomal integrant or as an extra chromosomal element.

"Promoter" as used herein refers to a nucleic acid sequence that regulates expression of a gene.

"Co-transfection" means the process of transfecting a eukaryotic cell with more than one exogenous gene foreign to the cell, one of which may confer a selectable phenotype on the cell.

The purified and isolated DNA sequence having protein production increasing activity also comprises, besides one or more bent DNA element, at least one binding site for a DNA binding protein.

Usually the DNA binding protein is a transcription factor. Examples of transcription factors are the group comprising the polyQpolyP domain proteins.
Another example of a transcription factor is a transcription factor selected from the group comprising SATB1, NMP4, MEF2, S8, DLX1, FREAC7, BRN2, GATA 1/3, TATA, Bright, MSX or a combination of two or more thereof. PolyQpolyP domain proteins and SATB1, NMP4, MEF2, S8, DLX1, FREAC7, BRN2, GATA 1/3, TATA, Bright, MSX or a combination of two or more of these transcription factors are preferred. Most preferred are SATB1, NMP4, MEF2 and polyQpolyP domain proteins.

SATB1, NMP4 and MEF2, for example, are known to regulate the development and/or tissue-specific gene expression in mammals. These transcription factors have the capacity to alter DNA geometry, and reciprocally, binding to DNA as an allosteric ligand modifies their structure. Recently, SATB1 was found to form a cage-like structure circumscribing heterochromatin (Cai, S., H.J. Han, and T. Kohwi-Shigematsu, "Tissue-

specific nuclear architecture and gene expression regulated by SATB1" *Nat Genet*, 2003. 34(1): p. 42-51).

A further aspect of the invention is to provide a method for identifying a MAR sequence using a Bioinformatic tool comprising the computing of values of one or more DNA sequence features corresponding to DNA bending, major groove depth and minor groove width potentials, melting temperature. Preferably, the identification of one or more DNA sequence features further comprises a further DNA sequence feature corresponding to binding sites for DNA binding proteins, which is also computed with this method.

The bioinformatic tool used for the present method is preferably, SMAR Scan, which contains algorithms developed by Gene Express (http://srs6.bionet.nsc.ru/srs6bin/cgi-bin/wgetz?-e+[FEATURES-SiteID:'nR']) and based on Levitsky *et al.*, 1999. These algorithms recognise profiles, based on dinucleotides weight-matrices, to compute the theoretical values for conformational and physicochemical properties of DNA.

Preferably, SMAR Scan uses the four theoretical criteria also designated as DNA sequence features corresponding to DNA bending, major groove depth and minor groove width potentials, melting temperature in all possible combination, using scanning windows of variable size (see Fig. 3). For each function used, a cut-off value has to be set. The program returns a hit every time the computed score of a given region is above the set cut-off value for all of the chosen criteria. Two data output modes are available to handle the hits, the first (called "profile-like") simply returns all hit positions on the query sequence and their corresponding values for the different criteria chosen. The second mode (called "contiguous hits ") returns only the positions of several contiguous hits and their corresponding sequence. For this mode, the minimum number of contiguous hits is another cut-off value that can be set, again with a tunable window size. This second mode is the default mode of SMAR Scan. Indeed, from a semantic point of view, a hit is considered as a core-unwinding element (CUE), and a cluster of CUEs accompanied by clusters of binding sites for relevant proteins is considered as a MAR. Thus, SMAR Scan considers only several contiguous hits as a potential MAR.

To tune the default cut-off values for the four theoretical structural criteria, experimentally validated MARs from SMARt DB (http://transfac.gbf.de/- SMARt DB)

were used. All the human MAR sequences from the database were retrieved and analyzed with SMAR Scan using the "profile-like" mode with the four criteria and with no set cut-off value. This allowed the setting of each function for every position of the sequences. The distribution for each criterion was then computed according to these data (see Fig. 1 and 3).

The default cut-off values of SMAR can for the bend, the major groove depth and the minor groove width were set at the average of the 75th quantile and the median. For the melting temperature, the default cut-off value should be set at the 75th quantile. The minimum length for the "contiguous-hits" mode should be set to 300 because it is assumed to be the minimum length of a MAR (see Fig. 8 and 9). However, one skilled in the art would be able to determine the cut-off values for the above-mentioned criteria for a given organism with minimal experimentation.

Preferably, DNA bending values are comprised between 3 to 5 ° (radial degree). Most preferably are situated between 3.8 to 4.4 °, corresponding to the smallest peak of Fig. 1.

Preferably the major groove depth values are comprised between 8.9 to 9.3 Å (Angström) and minor groove width values between 5.2 to 5.8 Å. Most preferably the major groove depth values are comprised between 9.0 to 9.2 Å and minor groove width values between 5.4 to 5.7 Å.

Preferably the melting temperature is situated between 55 to 75 ° C (Celsius degree). Most preferably the melting temperature is comprised between 55 to 62 ° C.

The DNA binding protein of which values can be computed by the method is usually a transcription factor preferably a polyQpolyP domain or a transcription factor selected from the group comprising SATB1, NMP4, MEF2, S8, DLX1, FREAC7, BRN2, GATA 1/3, TATA, Bright, MSX or a combination of two or more of these transcription factors. However, one skilled in the art would be able to determine other kinds of transcription factors in order to carry out the method according to the present invention.

The present invention also encompasses the use of purified and isolated DNA sequence comprising a first isolated matrix attachment region (MAR) nucleotide sequence which is a MAR nucleotide sequence selected from the group comprising the

sequences SEQ ID Nos 1 to 23, a sequence complementary thereof, a part thereof sharing at least 70% nucleotides in length, a molecular chimera thereof, a combination thereof and variants or a MAR nucleotide sequence of a cLysMAR element and/or fragment, a sequence complementary thereof, a part thereof sharing at least 70% nucleotides in length, a molecular chimera thereof, a combination thereof and variants for increasing protein production activity in a eukaryotic host cell.

Said purified and isolated DNA sequence usually further comprises one or more regulatory sequences, as known in the art e.g. a promoter and/or an enhancer, polyadenylation sites and splice junctions usually employed for the expression of the protein or may optionally encode a selectable marker. Preferably said purified and isolated DNA sequence comprises a promoter which is operably linked to a gene of interest.

The DNA sequences of this invention can be isolated according to standard PCR protocols and methods well known in the art.

Promoters which can be used provided that such promoters are compatible with the host cell are, for example, promoters obtained from the genomes of viruses such as polyoma virus, adenovirus (such as Adenovirus 2), papilloma virus (such as bovine papilloma virus), avian sarcoma virus, cytomegalovirus (such as murine or human cytomegalovirus immediate early promoter), a retrovirus, hepatitis-B virus, and Simian Virus 40 (such as SV 40 early and late promoters) or promoters obtained from heterologous mammalian promoters, such as the actin promoter or an immunoglobulin promoter or heat shock promoters. Such regulatory sequences direct constitutive expression.

Furthermore, the purified and isolated DNA sequence might further comprise regulatory sequences which are capable of directing expression of the nucleic acid preferentially in a particular cell type (e. g., tissue-specific regulatory elements are used to express the nucleic acid). Tissue-specific regulatory elements are known in the art. Non-limiting examples of suitable tissue-specific promoters include the albumin promoter (liver-specific; Pinkert,et al., 1987. Genes Dev.1: 268-277), lymphoid-specific promoters (Calame and Eaton, 1988. Adv. Immunol. 43: 235-275), in particular promoters of T cell receptors (Winoto and Baltimore, 1989. EMBOJ. 8: 729-733) and immunoglobulins

(Banerji, etal., 1983. Cell 33: 729-740; Queen and Baltimore, 1983. Cell 33:741-748), neuron-specific promoters (e. g., the neurofilament promoter;Byrne and Ruddle, 1989. Proc.Natl. Acad. Sci. USA 86: 5473-5477), pancreas-specific promoters (Edlund, et al., 1985. Science 230: 912-916), and mammary gland-specific promoters (e. g., milk whey promoter; U. S. Pat. No. 4,873,316 and European Application No. 264,166).

Developmentally-regulated promoters are also encompassed. Examples of such promoters include, e.g., the murine hox promoters (Kessel and Gruss, 1990. Science 249: 374-379) and thea-fetoprotein promoter (Campes and Tilghman, 1989. Genes Dev. 3: 537-546).

Regulatable gene expression promoters are well known in the art, and include, by way of non-limiting example, any promoter that modulates expression of a gene encoding a desired protein by binding an exogenous molecule, such as the CRE/LOX system, the TET system, the NFkappaB/UV light system, the Leu3p/isopropylmalate system, and theGLVPc/GAL4 system (See e. g., Sauer, 1998, Methods 14 (4): 381-92 ; Lewandoski, 2001, Nat. Rev. Genet 2 (10): 743-55; Legrand-Poels et al., 1998, J. Photochem. Photobiol. B. 45: 18; Guo et al., 1996, FEBS Lett. 390 (2): 191-5; Wang et al., PNAS USA, 1999,96 (15): 84838). However, one skilled in the art would be able to determine other kinds of promoters that are suitable in carrying out the present invention.

Enhancers can be optionally included the purified DNA sequence of the invention then belonging to the regulatory sequence, e.g. the promoter.

The gene of interest preferably encodes a protein (structural or regulatory protein). As used herein "protein" refers generally to peptides and polypeptides having more than about ten amino acids. The proteins may be "homologous" to the host (i.e., endogenous to the host cell being utilized), or "heterologous," (i.e., foreign to the host cell being utilized), such as a human protein produced by yeast. The protein may be produced as an insoluble aggregate or as a soluble protein in the periplasmic space or cytoplasm of the cell, or in the extracellular medium. Examples of proteins include hormones such as growth hormone, growth factors such as epidermal growth factor, analgesic substances like enkephalin, enzymes like chymotrypsin, and receptors to

hormones or growth factors and includes as well proteins usually used as a visualizing marker e.g. green fluorescent protein.

Preferably the purified DNA sequence further comprises at least a second isolated matrix attachment region (MAR) nucleotide sequence selected from the group comprising the sequences SEQ ID Nos 1 to 23 or cLysMAR, a sequence complementary thereof, a part thereof sharing at least 70% nucleotides in length, a molecular chimera thereof, a combination thereof and variants. The isolated matrix attachment region (MAR) nucleotide sequence might be identical or different. Alternatively, a first and a second identical MAR nucleotide sequence are used.

Preferably, the MAR nucleotide sequences are located at both the 5' and the 3' ends of the sequence containing the promoter and the gene of interest. But the invention also envisions the fact that said first and or at least second MAR nucleotide sequences are located on a sequence distinct from the one containing the promoter and the gene of interest.

Embraced by the scope of the present invention is also the purified and isolated DNA sequence comprising a first isolated matrix attachment region (MAR) nucleotide sequence which is a MAR nucleotide sequence selected from the group comprising the sequences SEQ ID Nos 1 to 23, a sequence complementary thereof, a part thereof sharing at least 70% nucleotides in length, a molecular chimera thereof, a combination thereof and variants or a MAR nucleotide sequence of a cLysMAR element and/or fragment, a sequence complementary thereof, a part thereof sharing at least 70% nucleotides in length, a molecular chimera thereof, a combination thereof and variants can be used for increasing protein production activity in a eukaryotic host cell by introducing the purified DNA sequence into a eukaryotic host cell according to well known protocols. Usually applied methods for introducing DNA into eukaryotic host cells usually applied are e.g. direct introduction of cloned DNA by microinjection or microparticle bombardment; use of viral vectors; encapsulation within a carrier system; and use of transfecting reagents such as calcium phosphate, diethylaminoethyl (DEAE) –dextran or commercial transfection systems like the Lipofect-AMINE 2000 (Invitrogen). Preferably, the transfection method used to introduce the purified DNA sequence into a eukaryotic host cell is the method for transfecting a eukaryotic cell as described below.

The purified and isolated DNA sequence is used in the form of a circular vector. Preferably, the purified and isolated DNA sequence is in the form of a linear DNA sequence as vector.

As used herein, "plasmid" and "vector" are used interchangeably, as the plasmid is the most commonly used vector form. However, the invention is intended to include such other forms of expression vectors, including, but not limited to, viral vectors (e. g., replication defective retroviruses, adenoviruses and adeno-associated viruses), which serve equivalent functions.

The present invention further encompasses a method for transfecting a eukaryotic host cell, said method comprising

a) introducing into said eukaryotic host cell at least one purified DNA sequence comprising at least one DNA sequence of interest and/or at least one purified and isolated DNA sequence comprising a MAR nucleotide sequence or other chromatin modifying elements,
b) subjecting within a defined time said transfected eukaryotic host cell to at least one additional transfection step with at least one purified DNA sequence comprising at least one DNA sequence of interest and/or with at least one purified and isolated DNA sequence comprising a MAR nucleotide sequence or other chromatin modifying elements
c) selecting said transfected eukaryotic host cell.

Preferably at least two up to four transfecting steps are applied in step b).

In order to select the successful transfected cells, a gene that encodes a selectable marker (e. g., resistance to antibiotics) is generally introduced into the host cells along with the gene of interest. The gene that encodes a selectable marker might be located on the purified DNA sequence comprising at least one DNA sequence of interest and/or at least one purified and isolated DNA sequence consisting of a MAR nucleotide sequence or other chromatin modifying elements or might optionally be co-introduced in separate form e.g. on a plasmid. Various selectable markers include those that confer resistance to drugs, such as G418, hygromycin and methotrexate. The amount of the drug can be adapted as desired in order to increase productivity

Usually, one or more selectable markers are used. Preferably, the selectable markers used in each distinct transfection steps are different. This allows selecting the transformed cells that are "multi-transformed" by using for example two different antibiotic selections.

Any eukaryotic host cell capable of protein production and lacking a cell wall can be used in the methods of the invention. Examples of useful mammalian host cell lines include human cells such as human embryonic kidney line (293 or 293 cells subcloned for growth in suspension culture, Graham et al., J. Gen Virol 36, 59 (1977)), human cervical carcinoma cells (HELA, ATCC CCL 2), human lung cells (W138, ATCC CCL 75), human liver cells (Hep G2, HB 8065); rodent cells such as baby hamster kidney cells (BHK, ATCC CCL 10), Chinese hamster ovary cells/-DHFR (CHO, Urlaub and Chasin, Proc. Natl. Acad. Sci. USA, 77, 4216 (1980)), mouse sertoli cells (TM4, Mather, Biol. Reprod 23, 243-251 (1980)), mouse mammary tumor (MMT 060562, ATCC CCL51); and cells from other mammals such as monkey kidney CV1 line transformed by SV40 (COS-7, ATCC CRL 1651); monkey kidney cells (CV1 ATCC CCL 70); African green monkey kidney cells (VERO-76, ATCC CRL-1587); canine kidney cells (MDCK, ATCC CCL 34); buffalo rat liver cells (BRL 3A, ATCC CRL 1442); myeloma (e.g. NS0) /hybridoma cells.
Preferred for uses herein are mammalian cells, more preferred are CHO cells.

The DNA sequence of interest of the purified and isolated DNA sequence is usually a gene of interest preferably encoding a protein operably linked to a promoter as described above. The purified and isolated DNA sequence comprising at least one DNA sequence of interest might comprise additionally to the DNA sequence of interest MAR nucleotide sequence or other chromatin modifying elements.

Purified and isolated DNA sequence comprising a MAR nucleotide sequence are for example selected from the group comprising the sequences SEQ ID Nos 1 to 23 and/or particular elements of the cLysMAR e.g. the B, K and F regions as well as fragment and elements and combinations thereof as described above. Other chromatin modifying elements are for example boundary elements (BEs), locus control regions (LCRs), and universal chromatin opening elements (UCOEs) (see Zahn-Zabal et al. already cited). An example of multiple transfections of host cells is shown in Example 12 (Table 3). The first transfecting step (primary transfection) is carried out with the gene of interest

(SV40EGFP) alone, with a MAR nucleotide sequence (MAR) alone or with the gene of interest and a MAR nucleotide sequence (MAR-SV40EGFP). The second transfecting step (secondary transfection) is carried out with the gene of interest (SV40EGFP) alone, with a MAR nucleotide sequence (MAR) alone or with the gene of interest and a MAR nucleotide sequence (MAR-SV40EGFP), in all possible combinations resulting from the first transfecting step.

Preferably the eukaryotic host cell is transfected by:

a) introducing a purified DNA sequence comprising one DNA sequence of interest and additionally a MAR nucleotide sequence,

b) subjecting within a defined time said transfected eukaryotic host cell to at least one additional transfection step with the same purified DNA sequence comprising one DNA sequence of interest and additionally a MAR nucleotide sequence of step a).

Surprisingly, a synergy between the first and second transfection has been observed. A particular synergy has been observed when MAR elements are present at one or both of the transfection steps. Multiple transfections of the cells with pMAR alone or in combination with various expression plasmids, using the method described above have been carried out. For example, Table 3 shows that transfecting the cells twice with the pMAR-SV40EGFP plasmid gave the highest expression of GFP and the highest degree of enhancement of all conditions (4.3 fold). In contrast, transfecting twice the vector without MAR gave little or no enhancement, 2.8-fold, instead of the expected two-fold increase. This proves that the presence of MAR elements at each transfection step is of particular interest to achieve the maximal protein synthesis.

As a particular example of the transfection method, said purified DNA sequence comprising at least one DNA sequence of interest can be introduced in form of multiple unlinked plasmids, comprising a gene of interest operably linked to a promoter, a selectable marker gene, and/or protein production increasing elements such as MAR sequences.

The ratio of the first and subsequent DNA sequences may be adapted as required for the use of specific cell types, and is routine experimentation to one ordinary skilled in the art.

The defined time for additional transformations of the primary transformed cells is tightly dependent on the cell cycle and on its duration. Usually the defined time corresponds to intervals related to the cell division cycle.
Therefore this precise timing may be adapted as required for the use of specific cell types, and is routine experimentation to one ordinary skilled in the art.
Preferably the defined time is the moment the host cell just has entered into the same phase of a second or a further cell division cycle, preferably the second cycle.
This time is usually situated between 6h and 48 h, preferably between 20h and 24h after the previous transfecting event.

The invention further comprises a transgenic organism wherein its genome has stably integrated at least one DNA sequence according to the present invention and/or that at least some of its cells have been transfected according to the above-described method.
Transgenic animals eukaryotic organisms which can be useful for the present invention are for example selected form the group comprising mammals (mouse, human, monkey etc) and in particular laboratory animals such as rodents in general, insects (drosophila, etc), fishes (zebra fish, etc.), amphibians (frogs, newt, etc..) and other simpler organisms such as c. elegans, yeast, etc....

Yet another object of the present invention is to provide a purified and isolated DNA sequence identified according to method of the present invention, having matrix attachment region (MAR) activity i.e. capable of having protein production increasing activity. Preferred purified and isolated DNA sequence identified according to present invention comprise a sequence selected from the sequences SEQ ID Nos 1 to 23 or a cLysMAR element and/or fragment, a sequence complementary thereof, a part thereof sharing at least 70% nucleotides in length, a molecular chimera thereof, a combination thereof and variants.

More preferably, the matrix attachment region (MAR) activity comprises a sequence selected from the sequences SEQ ID Nos 21 to 23, a sequence complementary thereof, a part thereof sharing at least 70% nucleotides in length, a molecular chimera thereof, a combination thereof and variants.

In the present invention, the cLysMAR element and/or fragment are consisting of at least one nucleotide sequence selected from the B, K and F regions.

A further object of the present invention is to provide a synthetic MAR sequence consisting of natural MAR element and/or fragments assembled between linker sequences.

Preferably, the synthetic MAR sequence is a cLysMAR and linker sequences are BglI-BamHI linker.

The present invention also provides for a cell transfection mixture or Kit comprising at least one purified and isolated DNA sequence according to the invention.

Also envisioned is a process for the production of a protein wherein a eukaryotic host cell is transfected according to the process as defined in the present invention and is cultured in a culture medium under conditions suitable for expression of the protein. Said protein is finally recovered according to any recovering process known to the skilled in the art.

Given as an example, the following process for protein production might be used. The eukaryotic host cell transfected with the transfection method of the present invention is used in a process for the production of a protein by culturing said cell under conditions suitable for expression of said protein and recovering said protein. Suitable culture conditions are those conventionally used for in vitro cultivation of eukaryotic cells as described e.g. in WO 96/39488. The protein can be isolated from the cell culture by conventional separation techniques such as e.g. fractionation on immunoaffinity or ion-exchange columns; precipitation; reverse phase HPLC; chromatography; chromatofocusing; SDS-PAGE; gel filtration One skilled in the art will appreciate that purification methods suitable for the polypeptide of interest may require modification to account for changes in the character of the polypeptide upon expression in recombinant cell culture.

The proteins that are produced according to this invention can be tested for functionality by a variety of methods. For example, the presence of antigenic epitopes and ability of the proteins to bind ligands can be determined by Western blot assays, fluorescence cell sorting assays, immunoprecipitation, immunochemical assays and/or

competitive binding assays, as well as any other assay which measures specific binding activity.

The proteins of this invention can be used in a number of practical applications including, but not limited to:

1. Immunization with recombinant host protein antigen as a viral/pathogen antagonist.
2. Production of membrane proteins for diagnostic or screening assays.
3. Production of membrane proteins for biochemical studies.
4. Production of membrane protein for structural studies.
5. Antigen production for generation of antibodies for immuno-histochemical mapping, including mapping of orphan receptors and ion channels.

The foregoing description will be more fully understood with reference to the following Examples. Such Examples, are, however, exemplary of methods of practising the present invention and are not intended to limit the scope of the invention.

EXAMPLES

Example 1: SMAR Scan and MAR sequences

A first rough evaluation of SMAR Scan was done by analyzing experimentally defined human MARs and non-MAR sequences. As MAR sequences, the previous results from the analysis of human MARs from SMARt Db were used to plot a density histogram for each criterion as shown in Fig. 1. Similarly, non-MAR sequences were also analyzed and plotted. As non-MAR sequences, all Ref-Seq-contigs from the chromosome 22 were used, considering that this latter was big enough to contain a negligible part of MAR sequences regarding the part of non-MAR sequences.

The density distributions shown in Fig. 1 are all skewed with a long tail. For the highest bend, the highest major groove depth and the highest minor groove width, the distributions are right skewed. For the lowest melting temperature, the distributions are left-skewed which is natural given the inverse correspondence of this criterion regarding the three others. For the MAR sequences, biphasic distributions with a second weak peak, are actually apparent. And between MAR and non-MAR sequences distributions, a clear shift is also visible in each plot.

Among all human MAR sequences used, in average only about 70% of them have a value greater than the 75th quantile of human MARs distribution, this for the four different criteria. Similarly concerning the second weak peak of each human MARs distribution, only 15% of the human MAR sequences are responsible of these outlying values. Among these 15% of human MAR sequences, most are very well documented MARs, used to insulate transgene from position effects, such as the interferon locus MAR, the beta-globin locus MAR (Ramezani A, Hawley TS, Hawley RG, Performance- and safety-enhanced lentiviral vectors containing the human interferon-beta scaffold attachment region and the chicken beta-globin insulator, *Blood*, 101:4717-4724, 2003), or the apolipoprotein MAR (Namciu, S, Blochinger KB, Fournier REK, Human matrix attachment regions in-sulate transgene expression from chromosomal position effects in Drosophila melanogaster, *Mol. Cell. Biol.*, 18:2382-2391, 1998).

Always with the same data, human MAR sequences were also used to determine the association between the four theoretical structural properties computed and the AT-

content. Fig. 2 represents the scatterplot and the corresponding correlation coefficient r for every pair of criteria.

Example 2: Distribution plots of MAR sequences by organism

MAR sequences from SMARt DB of other organisms were also retrieved and analyzed similarly as explained previously. The MAR sequences density distributions for the mouse, the chicken, the sorghum bicolor and the human are plotted jointly in Fig. 3.

Example 3: MAR prediction of the whole chromosome 22

All RefSeq contigs from the chromosome 22 were analyzed by SMAR Scan using the default settings this time. The result is that SMAR Scan predicted a total of 803 MARs, their average length being 446 bp, which means an average of one MAR predicted per 42 777 bp. The total length of the predicted MARs corresponds to 1% of the chromosome 22 length. The AT-content of the predicted regions ranged from 65,1% to 93.3%; the average AT-content of all these regions being 73.5%. Thus, predicted MARs were AT-rich, whereas chromosome 22 is not AT-rich (52.1% AT).

SMARTest was also used to analyze the whole chromosome 22 and obtained 1387 MAR candidates, their average length being 494 bp representing an average of one MAR predicted per 24 765 bp. The total length of the predicted MARs corresponds to 2% of the chromosome 22. Between all MARs predicted by the two softwares, 154 predicted MARs are found by both programs, which represents respectively 19% and 11% of SMAR Scan and SMARTest predicted MARs. Given predicted MARs mean length for SMAR Scan and SMARTest, the probability to have by chance an overlapping between SMAR Scan and SMARTest predictions is 0.0027% per prediction.

To evaluate the specificity of SMAR Scan predictions, SMAR Scan analyses were performed on randomly shuffled sequences of the chromosome 22 (Fig.4). Shuffled sequences were generated by three ways, by a segmentation of the chromosome 22 into non-overlapping windows of 10 bp and by separately shuffling the nucleotides in each window, by "scrambling" which means a permutation of all nucleotides of the chromosome and by "rubbling" which means a segmentation of the chromosome in fragments of 10 bp and then a random assembling of these fragments. The first shuffling method preserves the local nucleotide composition whereas the two other

methods destroy local information preserving only the global nucleotide composition. For each shuffling method, five shuffled chromosome 22 were generated and analyzed by SMAR Scan using the default settings. Concerning the number hits, an average of 3 519 170 hits (sd: 18 353) was found for the permutated chromosome 22 within non-overlapping windows of 10 bp, 171 936,4 hits (sd: 2 859,04) for the scrambled sequences and 24 708,2 hits (sd: 1 191,59) for the rubbled chromosome 22, which respectively represents 185% (sd: 1%), 9% (sd: 0.15%) and 1% (sd: 0.06 %) of the number of hits found with the native chromosome 22. For the number of MARs predicted, which thus means contiguous hits of length greater than 300, 1 997 MARs were predicted with the shuffled chromosome 22 within windows of 10 bp (sd: 31.2), only 2.4 MARs candidates were found in scrambled sequences (sd: 0.96) and none for the rubbled sequence, which respectively represents 249% and less than 0.3% of the number of predicted MARs found with the native chromosome 22.

Example 4:Accuracy of SMAR Scan prediction and comparison with other predictive tools

The accuracy of SMAR Scan was evaluated using six genomic sequences for which experimentally determined MARs have been mapped. In order to perform a comparison with other predictive tools, the sequences analyzed are the same with the sequences previously used to compare MAR-Finder and SMARTest. These genomic sequences are three plant and three human sequences (Table 1) totalizing 310 151 bp and 37 experimentally defined MARs. The results for SMARTest and MAR-Finder in Table 1 come from a previous comparison (Frisch M, Frech K, Klingenhoff A, Cartharius K, Liebich I and Werner T, In silico pre-diction of scaffold/matrix attachment regions in large genomic sequences, *Genome Research*, 12:349-354, 2001.).
MAR-Finder has been used with the default parameters excepted for the threshold that has been set to 0.4 and for the analysis of the protamine locus, the AT-richness rule has been excluded (to detect the non AT-rich MARs as was done for the protamine locus).

Sequence, description and reference	Length (kb)	Experimentally defined MARs positions (kb)	SMARTest prediction positions (kb)	MAR-Finder prediction positions (kb)	SMAR Scan prediction positions (kb)
Oryza Sativa putative ADP-glucose pyrophosphorylase subunit SH2 and putative NADPH dependant reductase A1 genes (U70541). [4]	30.034	0.0-1.2 5.4-7.4 17.3-18.5 20.0-23.1	- 6.5-7.0 15.2-15.7 16.2-16.6 17.6-18.3 19.6-20.1 20.7-21.3 23.6-23.9 25.0-25.4 27.5-27.9	- - 15.7-15.9 - 17.5-18.4 19.8-20.4 21.3-21.5 23.9-24.2 24.7-25.1 -	- - 15.6-16 - 17.6-18.2 21.8-22 - 23.4-23.8 - -
Sorghum bicolor ADP-glucose pyrophosphorylase subunit SH2. NADPH-dependant reducatse A1-b genes (AF010283). [4]	42.446	0.0-1.5 7.1-9.7 22.4-24.7 32.5-33.7 41.6-42.3	- - 21.3-21.9 22.9-24.0 - 27.3-27.6 - -	- - - 23.2-24.2 - 26.9-27.5 - -	- 7.4-7.7 21.5-21.8 22.9-23.2 23.6-24.0 27.3-27.6 33.4-33.9 -
Sorghum bicolor BAC clone 110K5 (AF124045). [37]	78.195	~0.9 ~5.8 ~6.3 ~9.3 ~15.0 ~18.5 ~21.9 ~23.3 ~25.6 ~29.1 ~34.6 ~44.1 ~48.5 ~57.9 ~62.9 ~67.1 ~69.3 ~73.7	- - - - 15.1-15.8 - 21.7-22.0 - - - - - 44.1-44.5 47.9-49.5 - - 63.1-63.7 - - 74.3-74.7	- - - - - - - - - - - - - 47.9-49.4 - - - - - -	- - - - - - 21.4-21.9 - - 29.2-29.5 - 39.0-40.0 - 48.1-48.6 48.8-49.3 - - - - 74.3-74.6
Human alpha-1-antitrypsin and corticosteroid binding globulin intergenic region (AF156545). [35]	30.461	2.6-6.3 22.0-30.4	5.5-6.0 - 25.7-26.2 27.5-27.8 - -	3.0-3.2 5.1-6.0 24.9-25.3 25.5-25.8 26.2-26.4 27.5-28.2	5.4-5.8 - 25.8-26.4 - - -
Human protamine locus (U15422). [24]	53.060	8.8-9.7 32.6-33.6 37.2-39.4 51.8-53.0	- - - -	8.0-8.9 33.9-34.8 33.9-34.8 -	- - - -
Human beta-globin locus (U01317). [21]	75.955	1.5-3.0 15.6-19.0 44.7-52.7 60.0-70.0	- 18.0-18.4 - 34.4-34.9 - 56.6-57.1 59.8-60.3 65.6-66.0	- 15.5-16.0 18.0-18.4 - 50.6-50.8 56.5-57.2 58.1-58.5 63.0-63.6	2.3-2.6 15.3-15.6 - - - - 62.8-63.1 -

			67.6-67.9 68.8-69.1	68.7-69.3 -	66.3-66.7 -
Sum(kb)	310.151	at least 56.1	14.5	13.8	9.5
Total numbers : Average kb /predicted MAR True positives [number of experimentally defined MAR found] False positives False negatives Specificity Sensitivity		37	28 11.076 19[14] 9 23 19/28= 68% 14/37= 38%	25 12.408 20[12] 5 25 20/25= 80% 12/37= 32%	22 14.097 17[14] 5 23 17/22= 77% 14/37= 38%

Table 1: Evaluation of SMAR Scan accuracy

Six different genomic sequences, three plant and three human sequences, for which experimentally defined MARs are known, were analyzed with MAR-Finder, SMARTest and SMAR Scan. True positive matches are printed in bold, minus (-) indicates false negative matches. Some of the longer experimentally defined MARs contained more than one in silico prediction, each of them was counted as true positive match. Therefore, the number of true in silico predictions is higher than the number of experimentally defined MARs found. Specificity is defined as the ratio of true positive predictions, whereas sensitivity is defined as the ratio of experimentally defined MARs found. * AT-rich rule excluded using MAR-Finder.

SMARTest predicted 28 regions as MARs, 19 (true positives) of these correlate with experimentally defined MARs (specificity: 68%) whereas 9 (32%) are locatedin non-MARs (false positives). As some of the longest experimentally determined MARs contains more than one in silico prediction, the 19 true positives correspond actually to 14 different experimentally defined MARs (sensitivity: 38%). MARFinder predicted 25 regions as MARs, 20 (specificity: 80%) of these correlate with experimentally defined MARs corresponding to 12 different experimentally defined MARs (sensitivity: 32%). SMAR Scan predicted 22 regions, 17 being true positives (specificity: 77%) matching 14 different experimentally defined MARs (sensitivity: 38%).

As another example, the same analysis has been applied to human chromosomes 1 and 2 and lead to the determination of 23 MARs sequences (SEQ ID N° 1 to 23). These sequences are listed in Annex 1 in ST25 format.

Example 5:Dissection of the chicken lysozyme gene 5'- MAR

The 3000 base pair 5'-MAR was dissected into smaller fragments that were monitored for effect on transgene expression in Chinese hamster ovary (CHO) cells. To do so, seven fragments of ~400 bp were generated by polymerase chain reaction (PCR). These PCR-amplified fragments were contiguous and cover the entire MAR sequence when placed end-to-end. Four copies of each of these fragments were ligated in a head-to-tail orientation, to obtain a length corresponding to approximately half of that of the natural MAR. The tetramers were inserted upstream of the SV40 promoter in pGEGFPControl, a modified version of the pGL3Control vector (Promega). The plasmid pGEGFPControl was created by exchanging the luciferase gene of pGL3Control for the EGFP gene from pEGFP-N1 (Clontech). The 5'-MAR-fragment-containing plasmids thus created were co-transfected with the resistance plasmid pSVneo in CHO-DG44 cells using LipofectAmine 2000 (Invitrogen) as transfection reagent, as performed previously (Zahn-Zabal, M., et al., "Development of stable cell lines for production or regulated expression using matrix attachment regions" *J Biotechnol*, 2001. 87(1): p. 29-42.). After selection of the antibiotic (G-418) resistant cells, polyclonal cell populations were analyzed by FACS for EGFP fluorescence.

Transgene expression was expressed at the percentile of high expressor cells, defined as the cells which fluorescence levels are at least 4 orders of magnitude higher than the average fluorescence of cells transfected with the pGEGFPControl vector without MAR. Fig. 5 shows that multimerized fragments B, K and F enhance transgene expression, despite their shorter size as compared to the original MAR sequence. In contrast, other fragments are poorly active or fully inactive.

Example 6:Specificity of B, K and F regions in the MAR context

The 5'-MAR was serially deleted from the 5'-end (Fig.6, upper part) or the 3'-end (Fig.6, lower part), respectively. The effect of the truncated elements was monitored in an assay similar to that described in the previous section. Figure 6 shows that the loss of ability to stimulate transgene expression in CHO cells was not evenly distributed.

In this deletion study, the loss of MAR activity coincided with discrete regions of transition which overlap with the 5'-MAR B-, K- and F-fragment, respectively. In 5' deletions, activity was mostly lost when fragment K and F were removed. 3' deletions that removed the F and b elements had the most pronounced effects. In contrast,

flanking regions A, D, E and G that have little or no ability to stimulate transgene expression on their own (Fig. 5), correspondingly did not contribute to the MAR activity in the 5'- and 3'-end deletion studies (Fig. 6).

Example 7:Structure of the F element

The 465 bp F fragment was further dissected into smaller sub-fragments of 234, 243, 213 bp and 122, 125 and 121 bp, respectively. Fragments of the former group were octamerized (8 copies) in a head-to-tail orientation, while those of the latter group were similarly hexa-decamerized (16 copies), to maintain a constant length of MAR sequence. These elements were cloned in pGEGFPControl vector and their effects were assayed in CHO cells as described previously. Interestingly, fragment FIII retained most of the activity of the full-length F fragment whereas fragment FII, which contains the right-hand side part of fragment FIII, lost all the ability to stimulate transgene expression (Fig. 7). This points to an active region comprised between nt 132 and nt 221 in the FIB fragment. Consistently, multiple copies of fragments FI and FIB, which encompass this region, displayed similar activity. FIIA on its own has no activity. However, when added to FIB, resulting in FIII, it enhances the activity of the former. Therefore FIIA appears to contain an auxiliary sequence that has little activity on its own, but that strengthens the activity of the minimal domain located in FIB.

Analysis of the distribution of individual motifs within the lysozyme gene 5'-MAR is shown in Fig. 8 A, along with some additional motifs that we added to the analysis. Most of these motifs were found to be dispersed throughout the MAR element, and not specifically associated with the active portions. For instance, the binding sites of transcription factors and other motifs that have been associated with MARs were not preferentially localized in the active regions. It has also been proposed that active MAR sequences may consist of combination of distinct motifs. Several computer programs (MAR Finder, MARscan, SMARTest, SIDD duplex stability) have been reported to identify MARs as regions of DNA that associate with the DNA matrix. They are usually based on algorithms that utilizes a predefined series of sequence-specific patterns that have previously been suggested as containing MAR activity, as exemplified by MAR Finder, now known as MAR Wiz. The output of these programs did not correlate well with the transcriptionally active portions of the *cLysMAR*. For instance, peaks of activity obtained with MAR Finder did not clearly match active MAR sub-portion, as for instance the B fragment is quite active in vivo but scores negative with MAR Finder (Fig. 8B,

compare the top and middle panels). Bent DNA structures, as predicted by this program, did not correlate well either with activity (Fig. 8B, compare the top and bottom panels). Similar results were obtained with the other available programs (data not shown).

The motifs identified by available MAR prediction computer methods are therefore unlikely to be the main determinants of the ability of the *cLysMAR* to increase gene expression. Therefore, a number of other computer tools were tested. Surprisingly, predicted nucleosome binding sequences and nucleosome disfavouring sequences were found to be arranged in repetitively interspersed clusters over the MAR, with the nucleosome favouring sites overlapping the active B, K and F regions. Nucleosome positioning sequences were proposed to consist of DNA stretches that can easily wrap around the nucleosomal histones, and they had not been previously associated with MAR sequences.

Nucleosome-favouring sequences may be modelled by a collection of DNA features that include moderately repeated sequences and other physico-chemical parameters that may allow the correct phasing and orientation of the DNA over the curved histone surface. Identification of many of these DNA properties may be computerized, and up to 38 different such properties have been used to predict potential nucleosome positions. Therefore, we set up to determine if specific components of nucleosome prediction programs might correlate with MAR activity, with the objective to construct a tool allowing the identification of novel and possibly more potent MARs from genomic sequences.

To determine whether any aspects of DNA primary sequence might distinguish the active B, K and F regions from the surrounding MAR sequence, we analyzed the 5'-MAR with MARScan. Of the 38 nucleosomal array prediction tools, three were found to correlate with the location of the active MAR sub-domains (Fig. 9A). Location of the MAR B, K and F regions coincides with maxima for DNA bending, major groove depth and minor groove width. A weaker correlation was also noted with minima of the DNA melting temperature, as determined by the GC content. Refined mapping over the MAR F fragment indicated that the melting temperature valley and DNA bending summit indeed correspond the FIB sub-fragment that contains the MAR minimal domain (Fig. 9B). Thus active MAR portions may correspond to regions predicted as curved DNA

regions by this program, and we will refer to these regions as CUE-B, CUE-K and CUE-F in the text below. Nevertheless, whether these regions correspond to actual bent DNA and base-pair unwinding regions is unknown, as they do not correspond to bent DNA as predicted by MAR Wiz (Fig.9B).

Example 8:Imprints of other regulatory elements in the F fragment

Nucleosome positioning features may be considered as one of the many specific chromatin codes contained in genomic DNA. Although this particular code may contribute to the activity of the F region, it is unlikely to determine MAR activity alone, as the 3' part of the F region enhanced activity of the minimal MAR domain contained in the FIB portion. Using the MatInspector program (Genomatix), we searched for transcription factor binding sites with scores higher than 0.92 and found DNA binding sequences for the NMP4 and MEF2 proteins in the 3' part of the F fragment (Fig. 8B). To determine whether any of these transcription factor-binding sites might localize close to the B and K active regions, the entire 5'-MAR sequence was analyzed for binding by NMP4 and MEF2 and proteins reported to bind to single-stranded or double-stranded form of BURs. Among those, SATB1 (special AT-rich binding protein 1) belongs to a class of DNA-binding transcription factor that can either activate or repress the expression of nearby genes. This study indicated that specific proteins such as SATB1, NMP4 (nuclear matrix protein 4) and MEF2 (myogenic enhancer factor 2), have a specific distribution and form a framework around the minimal MAR domains of *cLysMAR* (Fig. 10). The occurrence of several of these NMP4 and SATB1 binding sites has been confirmed experimentally by the EMSA analysis of purified recombinant proteins (data not shown).

Example 9:Construction of artificial MARs by combining defined genetic elements

To further assess the relative roles of the various MAR components, the *cLysMAR* was deleted of all three CUE regions (Fig. 11, middle part), which resulted in the loss of part of its activity when compared to the complete MAR sequence similarly assembled from all of its components as a control (Fig. 11, top part). Consistently, one copy of each CUE alone, or one copy of each of the three CUEs assembled head-to-tail, had little activity in the absence of the flanking sequences. These results strengthen the conclusion that optimal transcriptional activity requires the combination of CUEs with of

flanking sequences. Interestingly, the complete MAR sequence generated from each of its components, but containing also BglII-BamHI linker sequences (AGATCC) used to assemble each DNA fragment, displayed high transcriptional activity (6 fold activation) as compared to the 4.8 fold noted for the original MAR element in this series of assays (see Fig. 5).

We next investigated whether the potentially curved DNA regions may also be active in an environment different from that found in their natural MAR context. Therefore, we set up to swap the CUE-F, CUE-B and CUE-K elements, keeping the flanking sequences unchanged. The sequences flanking the CUE-F element were amplified by PCR and assembled to bracket the various CUEs, keeping their original orientation and distance, or without a CUE. These engineered ~1.8 kb MARs were then assayed for their ability to enhance transgene expression as above. All three CUE were active in this context, and therefore there action is not restricted to one given set of flanking sequences. Interestingly, the CUE-K element was even more active than CUE-F when inserted between the CUE-F flanking sequences, and the former composite construct exhibited an activity as high as that observed for the complete natural MAR (4.8 fold activation). What distinguishes the CUE-K element from CUE-F and CUE-B is the presence of overlapping binding sites for the MEF-2 and SatB1 proteins, in addition to its CUE feature. Therefore, fusing CUE-B with CUE-F-flanking domain results in a higher density of all three binding sites, which is likely explanation to the increased activity. These results indicate that assemblies of CUEs with sequences containing binding sites for proteins such as NMP4, MEF-2, SatB1, and/or polyPpolyQ proteins constitute potent artificial MAR sequences.

Example 10:Expression vectors

Three expression vectors according to the present invention are represented on Figure 12.

Plasmid pPAG01 is a 5640 bp pUC19 derivative. It contains a 2980 bp chicken DNA fragment cloned in *BamH1* and *XbaI* restriction sites. The insert comes from the border of the 5'-end of the chicken lyzozyme locus and has a high A/T-content.

Plasmid pGEGFP (also named pSV40EGFP) control is a derivative of the pGL3-control vector (Promega) in which the luciferase gene sequence has been replaced by

the EGFP gene sequence form the pEGFP-N1 vector (Clontech). The size of pGEGFP plasmid is 4334bp.

Plasmid pUbCEGFP control is a derivative of the pGL3 wit an Ubiquitin promoter.

Plasmid pPAG01GFP (also named pMAR-SV40EGFP) is a derivative of pGEGFP with the 5'-Lys MAR element cloned in the MCS located just upstream of the SV40 promoter. The size of the pPAG01EGF plasmid is 7285bp.

Example 11:Effect of the additional transfection of primary transfectant cells on transgene expression

One day before transfection, cells were plated in a 24-well plate, in growth medium at a density of 1.35 x 10^5 cells/well for CHO-DG44 cells. 16 hours post-inoculum, cells were transfected when they reached 30-40% confluence, using Lipofect-AMINE 2000 (hereinafter LF2000), according to the manufacturer's instructions (Invitrogen). Twenty-seven microliters of serum free medium (Opti-MEM; Invitrogen) containing 1.4 µl of LF2000 were mixed with 27 µl of Opti-MEM containing 830 ng of linear plasmid DNA. The antibiotic selection plasmid (pSVneo) amounted to one tenth of the reporter plasmid bearing the GFP transgene. The mix was incubated at room temperature for 20 min. to allow the DNA-LF2000 complexes to form. The mixture was diluted with 300 µl of Opti-MEM and poured into previously emptied cell-containing wells. Following 3 hours incubation of the cells with the DNA mix at 37°C in a CO_2 incubator, one ml of DMEM-based medium was added to each well. The cells were further incubated for 24 hours in a CO_2 incubator at 37°C. The cells were then transfected a second time according to the method described above, except that the resistance plasmid carried another resistance gene (pSVpuro). Twenty-four hours after the second transfection, cells were passaged and expanded into a T-75 flask containing selection medium supplemented with 500 µg/ml G-418 and 5 µg/ml puromycin. After a two week selection period, stably transfected cells were cultured in 6-well plates. Alternatively, the cell population was transfected again using the same method, but pTKhygro (Clontech) and pSVdhfr as resistance plasmids. The expression of GFP was analysed with Fluorescence-activated cell sorter (FACS) and with a Fluoroscan.

Fig.13 shows that the phenotype of the twice-transfected cells (hereafter called secondary transfectants) not only was strongly coloured, such that special bulb and

filter were not required to visualize the green color from the GFP protein, but also contained a majority of producing cells (bottom right-hand side FACS histogram) as compared to the parental population (central histogram). This level of fluorescence corresponds to specific cellular productivities of at least 10 pg per cell per day. Indeed, cells transfected only one time (primary transfectants) that did not express the marker protein were almost totally absent from the cell population after re-transfection. Bars below 10^1 units of GFP fluorescence amounted 30% in the central histogram and less than 5% in the right histogram. This suggested that additional cells had been transfected and successfully expressed GFP.

Strikingly, the amount of fluorescence exhibited by re-transfected cells suggested that the subpopulation of cells having incorporated DNA twice expressed much more GFP than the expected two-fold increase. Indeed, the results shown in Table 2 indicate that the secondary transfectants exhibited, on average, more than the two-fold increase of GFP expected if two sets of sequences, one at each successive transfection, would have been integrated independently and with similar efficiencies. Interestingly, this was not dependent on the promoter sequence driving the reporter gene as both viral and cellular promoter-containing vectors gave a similar GFP enhancement (compare lane 1 and 2). However, the effect was particularly marked for the MAR-containing vector as compared to plasmids without MAR- (lane 3), where the two consecutive transfections resulted in a 5.3 and 4.6 fold increase in expression, in two distinct experiments.

Type of plasmids	Primary transfection	Secondary transfection	EGFP fluorescence Fold increase
pUbCEGFP	4'992	14'334	2.8
pSV40EGFP	4'324	12'237	2.8
pMAR-SV40EGFP	6'996	36'748	5.3

Type of plasmids	Primary transfection	Secondary transfection	EGFP fluorescence Fold increase
pUbCEGFP	6'452	15'794	2.5
pSV40EGFP	4'433	11'735	2.6
pMAR-SV40EGFP	8'116	37'475	4.6

Table 2. Effect of re-transfecting primary transfectants at 24 hours interval on GFP expression. Two independent experiments are shown. The resistance plasmid pSVneo was co-transfected with various GFP expression vectors. One day post-transfection, cells were re-transfected with the same plasmids with the difference

that the resistance plasmid was changed for pSVpuro. Cells carrying both resistance genes were selected on 500 µg/ml G-418 and 5µg/ml puromycin and the expression of the reporter gene marker was quantified by Fluoroscan. The fold increases correspond to the ratio of fluorescence obtained from two consecutive transfections as compared to the sum of fluorescence obtained from the corresponding independent transfections. The fold increases that were judged significantly higher are shown in bold, and correspond to fluorescence values that are consistently over 2-fold higher than the addition of those obtained from the independent transfections.

The increase in the level of GFP expression in multiply tranfected cells was not expected from current knowledge, and this effect had not been observed previously.

Taken together, the data presented here support the idea that the plasmid sequences that primarily integrated into the host genome would facilitate integration of other plasmids by homologous recombination with the second incoming set of plasmid molecules. Plasmid recombination events occur within a 1-h interval after the plasmid DNA has reached the nucleus and the frequency of homologous recombination between co-injected plasmid molecules in cultured mammalian cells has been shown to be extremely high, approaching unity (Folger, K.R., K. Thomas, and M.R. Capecchi, Nonreciprocal exchanges of information between DNA duplexes coinjected into mammalian cell nuclei. Mol Cell Biol, 1985. 5(1): p. 59-69], explaining the integration of multiple plasmid copies. However, homologous recombination between newly introduced DNA and its chromosomal homolog normally occurs very rarely, at a frequency of 1 in 10^3 cells receiving DNA to the most [Thomas, K.R., K.R. Folger, and M.R. Capecchi, High frequency targeting of genes to specific sites in the mammalian genome. Cell, 1986. 44(3): p. 419-28.]. Thus, the results might indicate that the MAR element surprisingly acts to promote such recombination events. MARs would not only modify the organization of genes in vivo, and possibly also allow DNA replication in conjunction with viral DNA sequences, but they may also act as DNA recombination signals.

Example 12:MARs mediate the unexpectedly high levels of expression in multiply transfected cells

If MAR-driven recombination events were to occur in the multiple transfections process, we expect that the synergy between the primary and secondary plasmid DNA would be

affected by the presence of MAR elements at one or both of the transfection steps. We examined this possibility by multiply transfections of the cells with pMAR alone or in combination with various expression plasmids, using the method described previously. Table 3 shows that transfecting the cells twice with the pMAR-SV40EGFP plasmid gave the highest expression of GFP and the highest degree of enhancement of all conditions (4.3 fold). In contrast, transfecting twice the vector without MAR gave little or no enhancement, 2.8-fold, instead of the expected two-fold increase. We conclude that the presence of MAR elements at each transfection step is necessary to achieve the maximal protein synthesis.

Table 3

Primary transfection		Secondary transfection		
Type of plasmid	EGFP-fluorescence	Type of plasmid	EGFP-fluorescence	Fold increase
pMAR	0	pMAR pSV40EGFP pMAR-SV40EGFP	0 15'437 30'488	0 2.3-2.5 2.6-2.7
pMAR-SV40EGFP	11'278	pMAR-SV40EGFP pMAR	47'027 12'319	4.3-5.3 1.0-1.1
pSV40EGFP	6'114	pSV40EGFP pMAR	17'200 11'169	2.8 1.8-2.3

Interestingly, when cells were first transfected with pMAR alone, and then re-transfected with pSV40EGFP or pMAR-SV40EGFP, the GFP levels were more than doubled as compared to those resulting from the single transfection of the later plasmids (2.5 and 2.7 fold respectively, instead of the expected 1-fold). This indicates that the prior transfection of the MAR can increase the expression of the plasmid used in the second transfection procedure. Because MARs act only locally on chromatin structure and gene expression, this implies that the two types of DNA may have integrated at a similar chromosomal locus. In contrast, transfecting the GFP expression vectors alone, followed by the MAR element in the second step, yielded little or no improvement of the GFP levels. This indicates that the order of plasmid transfection is important, and that the first transfection event should contain a MAR element to allow significantly higher levels of transgene expression.

If MAR elements favoured the homologous recombination of the plasmids remaining in episomal forms from the first and second transfection procedures, followed by their co-

integration at one chromosomal locus, one would expect that the order of plasmid transfection would not affect GFP levels. However, the above findings indicate that it is more favourable to transfect the MAR element in the first rather than in the second transfection event. This suggests the following molecular mechanism: during the first transfection procedure, the MAR elements may concatemerize and integrate, at least in part, in the cellular chromosome. This integrated MAR DNA may in turn favour the further integration of more plasmids, during the second transfection procedure, at the same or at a nearby chromosomal locus.

Example 13:MARs as long term DNA transfer facilitators

If integrated MARs mediated a persistent recombination-permissive chromosomal structure, one would expect high levels of expression even if the second transfection was performed long after the first one, at a time when most of the transiently introduced episomal DNA has been eliminated. To address this possibility, the cells from Table 3, selected for antibiotic resistance for three weeks, were transfected again once or twice and selected for the incorporation of additional DNA resistance markers. The tertiary, or the tertiary and quaternary transfection cycles, were performed with combinations of pMAR or pMAR-SV40EGFP, and analyzed for GFP expression as before.

Table 4

Tertiary transfection			Quaternary transfection		
Type of plasmid	EGFP-fluorescence	Fold increase	Type of plasmid	EGFP-fluorescence	Fold increase
pMAR	18368	2.2	pMAR pMAR-SV40EGFP	43'186 140'000	2.4 7.6
pMAR-SV40EGF	16544	2.0	pMAR-SV40EGFP pMAR	91'000 33'814	5.5 2.0

Table 4. MARs act as facilitator of DNA integration. The pMAR-SV40EGFP/ pMAR-SV40EGFP secondary transfectants were used in a third cycle of transfection at the end of the selection process. The tertiary transfection was accomplished with pMAR or pMAR-SV40EGFP, and pTKhygro as selection plasmid, to give tertiary transfectants. After 24 hours, cells were transfected again with either plasmid and pSVdhfr, resulting

in the quaternary transfectants which were selected in growth medium containing 500 µg/ml G-418 and 5µg/ml puromycin, 300 µg/ml hygromycin B and 5µM methotrexate. The secondary transfectants initially exhibited a GFP fluorescence of 8300. The fold increases correspond to the ratio of fluorescence obtained from two consecutive transfections as compared to the sum of fluorescence obtained from the corresponding independent transfections. The fold increases that were judged significantly higher are shown in bold, and correspond to fluorescence values that are 2-fold higher than the addition of those obtained from the independent transfections.

These results show that loading more copies of pMAR or pMAR-SV40EGFP resulted in similar 2-fold enhancements of total cell fluorescence. Loading even more of the MAR in the quaternary transfection further enhanced this activity by another 2.4-fold. This is consistent with our hypothesis that newly introduced MAR sequences may integrate at the chromosomal transgene locus by homologous recombination and thereby further increase transgene expression.

When the cells were transfected a third and fourth time with the pMAR-SV40EGFP plasmid, GFP activity further increased, once again to levels not expected from the addition of the fluorescence levels obtained from independent transfections. GFP expression reached levels that resulted in cells visibly glowing green in day light (Fig.14). These results further indicate that the efficiency of the quaternary transfection was much higher than that expected from the efficacy of the third DNA transfer, indicating that proper timing between transfections is crucial to obtain the optimal gene expression increase, one day being preferred over a three weeks period.

We believe that MAR elements favour secondary integration events in increasing recombination frequency at their site of chromosomal integration by relaxing closed chromatin structure, as they mediate a local increase of histone acetylation (Yasui, D., et al., SATB1 targets chromatin remodelling to regulate genes over long distances. *Nature*, 2002. 419(6907): p. 641-5.]. Alternatively, or concomitantly, MARs potentially relocate nearby genes to subnuclear locations thought to be enriched in trans-acting factors, including proteins that can participate in recombination events such as topoisomerases. This can result in a locus in which the MAR sequences can bracket the pSV40EGFP repeats, efficiently shielding the transgenes from chromatin-mediated silencing effects.

CLAIMS

1. A purified and isolated DNA sequence having protein production increasing activity characterized in that said DNA sequence comprises
   a) at least one bent DNA element,
   b) and at least one binding site for a DNA binding protein.

2. The purified and isolated DNA sequence of claim 1, characterized in that said bent DNA element is a MAR nucleotide sequence selected from the group comprising the sequences SEQ ID Nos 1 to 23, a sequence complementary thereof, a part thereof sharing at least 70% nucleotides in length, a molecular chimera thereof, a combination thereof and variants.

3. The purified and isolated DNA sequence of claim 1, characterized in that said bent DNA element is a cLysMAR element and/or fragment, a sequence complementary thereof, a part thereof sharing at least 70% nucleotides in length, a molecular chimera thereof, a combination thereof and variants.

4. The purified and isolated DNA sequence of claim 3, characterized in that said part thereof is a nucleotide sequence selected from the B, K and F regions.

5. The purified and isolated sequence of claims 1 to 4, characterized in that said DNA binding protein is a transcription factor.

6. The purified and isolated sequence of claim 5, characterized in that the transcription factor is selected from the group comprising the polyQpolyP domain proteins.

7. The purified and isolated sequence of claim 5, characterized in that the transcription factor is selected from the group comprising SATB1, NMP4, MEF2, S8, DLX1, FREAC7, BRN2, GATA 1/3, TATA, Bright, MSX or a combination of two or more of these transcription factors.

8. A method for identifying a MAR sequence using a Bioinformatic tool comprising the computing of values of one or more DNA sequence features corresponding to DNA bending, major groove depth and minor groove width potentials, melting temperature.

9. The method according to claim 8, characterized in that said Bioinformatic tool contains algorithms recognising profiles, based on dinucleotides weight-matrices, to compute values for one or more of said DNA sequence features corresponding to DNA bending, major groove depth and minor groove width potentials, and melting temperature.

10. The method according to claim 9, characterized in that said Bioinformatic tool computes values for all of said DNA sequence features.

11. The method according to claim 10, characterized in that said Bioinformatic tool is SMAR Scan.

12. The method according to claim 8-11, characterized in that the identification of one or more DNA sequence features further comprises a feature corresponding to one or more binding sites for DNA binding proteins.

13. The method according to claim 12, characterized in that said DNA binding protein is a transcription factor.

14. The method according to claim 13, characterized in that the transcription factor is selected from the group comprising polyQpolyP domain proteins or transcription factors.

15. The method according to claim 12, characterized in that the DNA binding protein is selected from the group comprising SATB1, NMP4, MEF2, S8, DLX1, FREAC7, BRN2, GATA 1/3, TATA, Bright, MSX or a combination of two or more of these transcription factors.

16. The method according to claims 8-15, characterized in that values for the identification of DNA bending are comprised between 3 to 5 °.

17. The method according to claim 16, characterized in that values for the identification of DNA bending are comprised between 3.8 to 4.4 °.

18. The method according to claims 8-17 characterized in that values for the identification of the major groove depth are comprised between 8.9 to 9.3 Å and values for the identification of minor groove width are comprised between 5.2 to 5.8 Å.

19. The method according to claims 18, characterized in that values for the identification of major groove depth are comprised between 9.0 to 9.3 Å and values for the identification of minor groove width are comprised between 5.4 to 5.7 Å.

20. The method according to claims 8-19, characterized in that the melting temperature is comprised between 55 to 75 ° C.

21. The method according to claim 20, characterized in that the melting temperature is comprised between 55 to 62 ° C.

22. The use of a purified and isolated DNA sequence comprising a first isolated matrix attachment region (MAR) nucleotide sequence which is a MAR nucleotide sequence selected from the group comprising the sequences SEQ ID Nos 1 to 23 and a cLysMAR element and/or fragment, a sequence complementary thereof, a part thereof sharing at least 70% nucleotides in length, a molecular chimera thereof, a combination thereof and variants for increasing protein production activity in a eukaryotic host cell.

23. The use of the purified and isolated DNA sequence of claim 22, characterized in that said purified and isolated DNA sequence further comprises a promoter operably linked to a gene of interest.

24. The use of the purified and isolated DNA sequence of claims 22-23, characterized in that said purified and isolated DNA sequence further comprises at least a second isolated matrix attachment region (MAR) nucleotide sequence which is a MAR nucleotide sequence selected from the group comprising the sequences SEQ ID Nos 1 to 23 and a cLysMAR element and/or fragment, a sequence complementary thereof, a part thereof sharing at least 70% nucleotides in length, a molecular chimera

thereof, a combination thereof and variants for increasing protein production activity in a eukaryotic host cell.

25. The use of the purified and isolated DNA sequence of claim 24, characterized in that said first and at least second MAR sequences are located at both the 5' and the 3' ends of the sequence containing the promoter and the gene of interest.

26. The use of the purified and isolated DNA sequence of claim 24, characterized in that said first and or at least second MAR sequences are located on a sequence distinct from the one containing the promoter and the gene of interest.

27. The use of the purified and isolated DNA sequence of any of claims 22-26, characterized in that said purified DNA sequence is in the form of a linear DNA sequence as vector.

28. A method for transfecting a eukaryotic host cell, said method comprising

a) introducing into said eukaryotic host cell at least one purified DNA sequence comprising at least one DNA sequence of interest and/or at least one purified and isolated DNA sequence consisting of a MAR nucleotide sequence or other chromatin modifying elements,
b) subjecting within a defined time said transfected eukaryotic host cell to at least one additional transfection step with at least one purified DNA sequence comprising at least one DNA sequence of interest and/or with at least one purified and isolated DNA sequence consisting of a MAR nucleotide sequence or other chromatin modifying elements
c) selecting said transfected eukaryotic host cell.

29. The method of claim 28, characterized in that said DNA sequence of interest is a gene of interest coding for a protein operably linked to a promoter.

30. The method of claim 29, characterized in that the selected transfected eukaryotic host cells are high protein producer cells with a production rate of at least 10 pg per cell per day.

31. The method of claims 28-30, characterized in that the MAR nucleotide is selected from the group comprising the sequences SEQ ID Nos 1 to 23 and a cLysMAR element, and/or fragment, a sequence complementary thereof, a part thereof sharing at least 70% nucleotides in length, a molecular chimera thereof, a combination thereof and variants.

32. The method of claims 28-31, characterized in that the defined time corresponds to intervals related to the cell division cycle.

33. The method of claim 32, characterized in that the defined time is the moment the host cell just has entered into a second cell division cycle.

34. A purified and isolated DNA sequence identified according to claims 8 to 21.

35. The purified and isolated DNA sequence of claim 34, having matrix attachment region (MAR) activity comprising a sequence selected from the sequences SEQ ID Nos 1 to 23, a sequence complementary thereof, a part thereof sharing at least 70% nucleotides in length, a molecular chimera thereof, a combination thereof and variants.

36. The purified and isolated DNA sequence of claim 35, having matrix attachment region (MAR) activity comprising a sequence selected from the sequences SEQ ID Nos 21 to 23, a sequence complementary thereof, a part thereof sharing at least 70% nucleotides in length, a molecular chimera thereof, a combination thereof and variants.

37. A purified and isolated cLysMAR element and/or fragment having protein production increasing activity, a sequence complementary thereof, a part thereof sharing at least 70% nucleotides in length, a molecular chimera thereof, a combination thereof and variants.

38. The cLysMAR element and/or fragment of claim 37 consisting of at least one nucleotide sequence selected from the B, K and F regions.

39. A synthetic MAR sequence consisting of natural MAR element and/or fragments assembled between linker sequences.

40. The synthetic MAR sequence of claim 39, characterized in that the MAR is cLysMAR.

41. The complete matrix attachment region of claim 39 to 40, characterized in that the linker sequences are BglI-BamHI linker.

42. A process for the production of a protein wherein
a) a eukaryotic host cell transfected according to claim 28 to 33, is cultured in a culture medium under conditions suitable for expression of said protein and
b) said protein is recovered.

43. A eukaryotic host cell transfected according to any one of claims 28 to 33.

44. A cell transfection mixture or kit comprising at least one purified and isolated DNA sequence according to claims 1 to 7 and/or 34 to 38.

45. A transgenic organism characterized in that its genome has stably integrated at least one DNA sequence according to claims 1 to 7 and 34 to 38, and/or that at least some of its cells have been transfected according to the method of claims 28-33.

46. The use of the purified and isolated DNA sequences of claim 34 as MAR sequences having protein production increasing activity.

ABSTRACT

The present invention relates to purified and isolated DNA sequences having protein production increasing activity and more specifically to the use of matrix attachment regions (MARs) for increasing protein production activity in a eukaryotic cell. Also disclosed is a method for the identification of said active regions, in particular MAR nucleotide sequences, and the use of these characterized active MAR sequences in a new multiple transfection method.

FIG.1

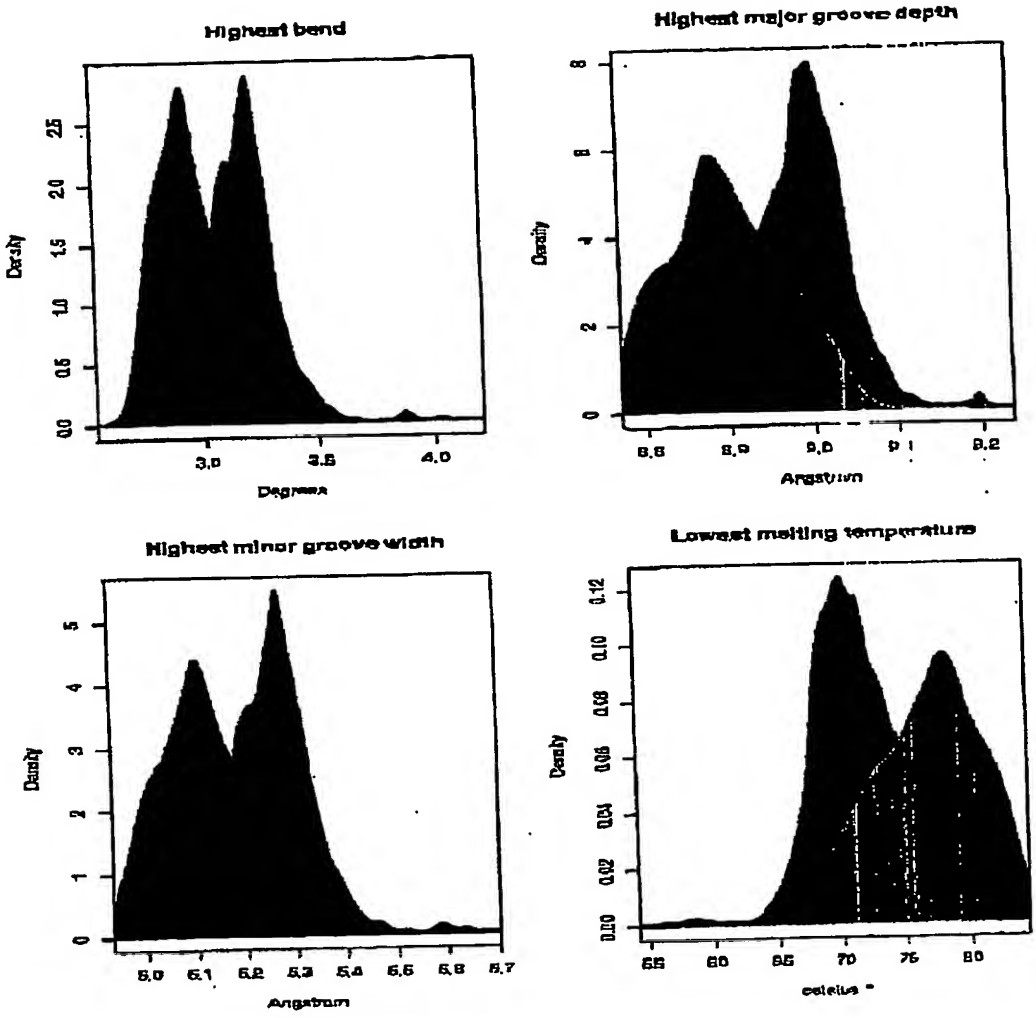

Highest bend
Density
Degrees
Highest major groove depth
Density
Angstrom
Highest minor groove width
Density
Angstrom
Lowest melting temperature
Density
celcius

FIG.2

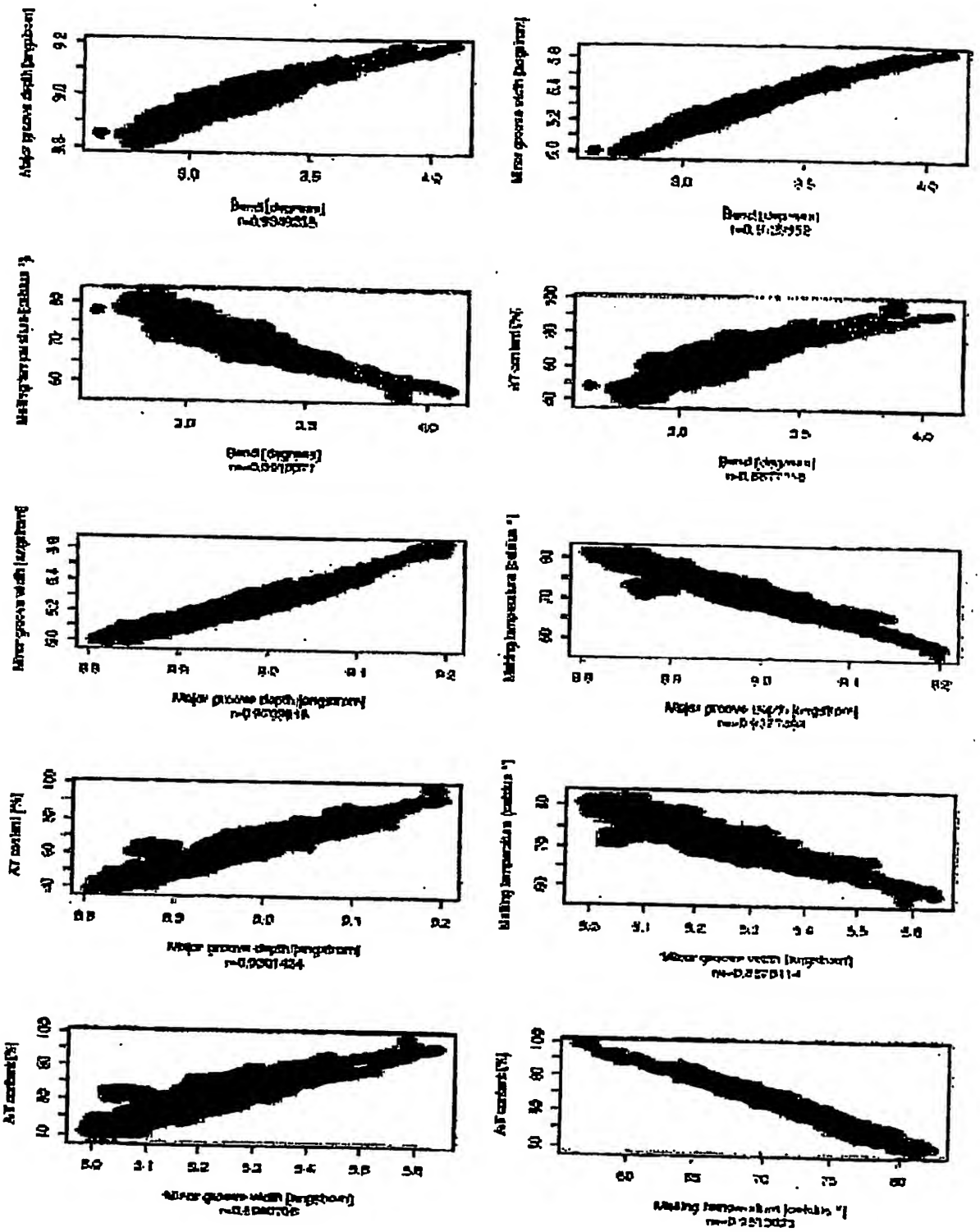


FIG.3

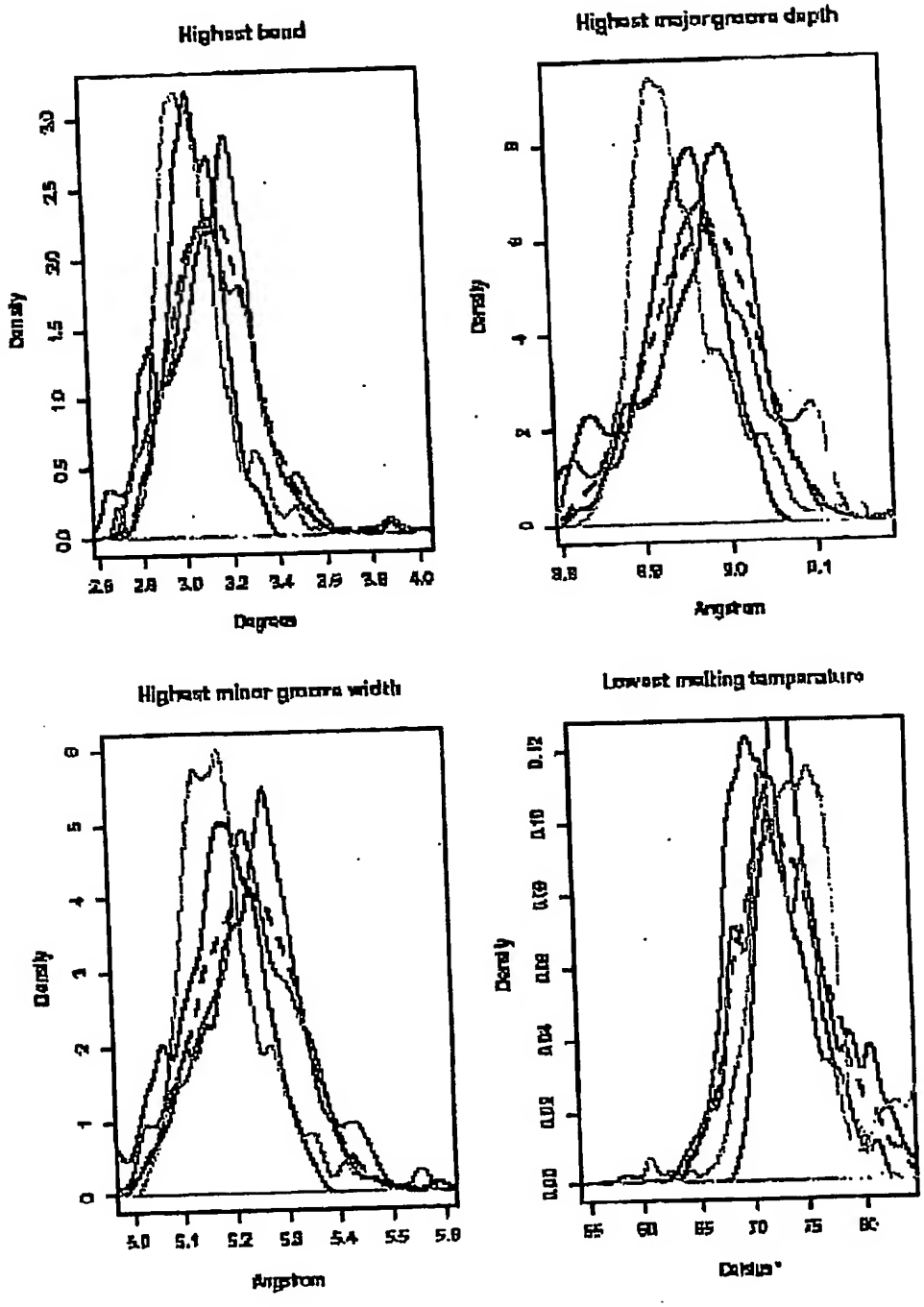

Highest bend
Density
Degrees
Highest majorgroove depth
Density
Angstrom
Highest minor groove width
Density
Angstrom
Lowest melting temperature
Density
Celsius

FIG.4

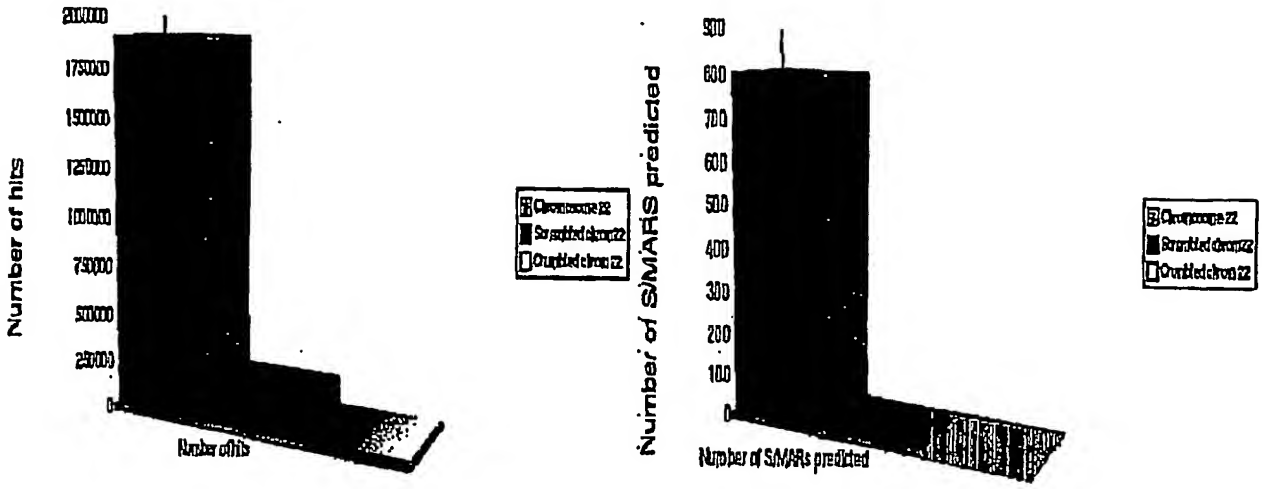

Number of hits
Number of hits
Number of SMARS predicted
Number of SMARS predicted

FIG.5

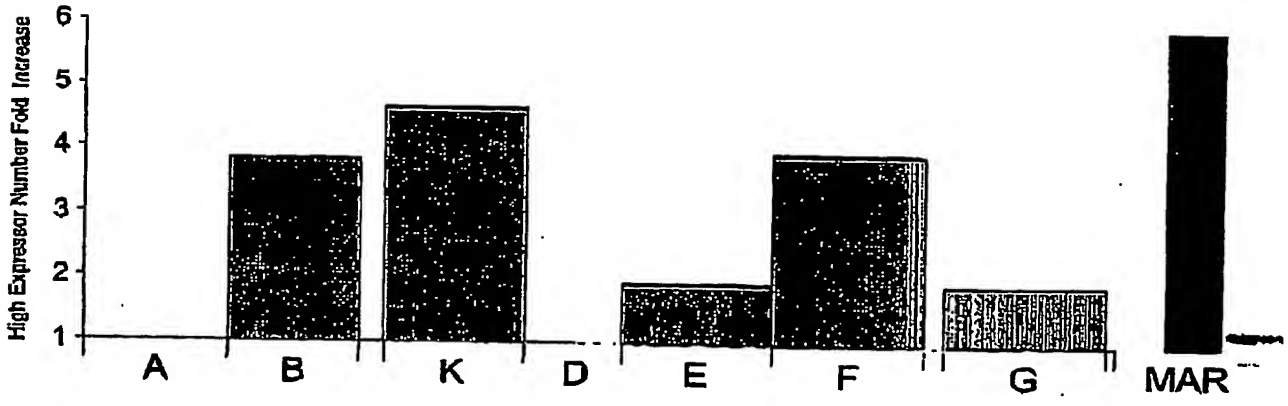

High Expressor Number Fold Increase
6
5
4
3
2
1
A
B
K
D
E
F
G
MAR

FIG.6

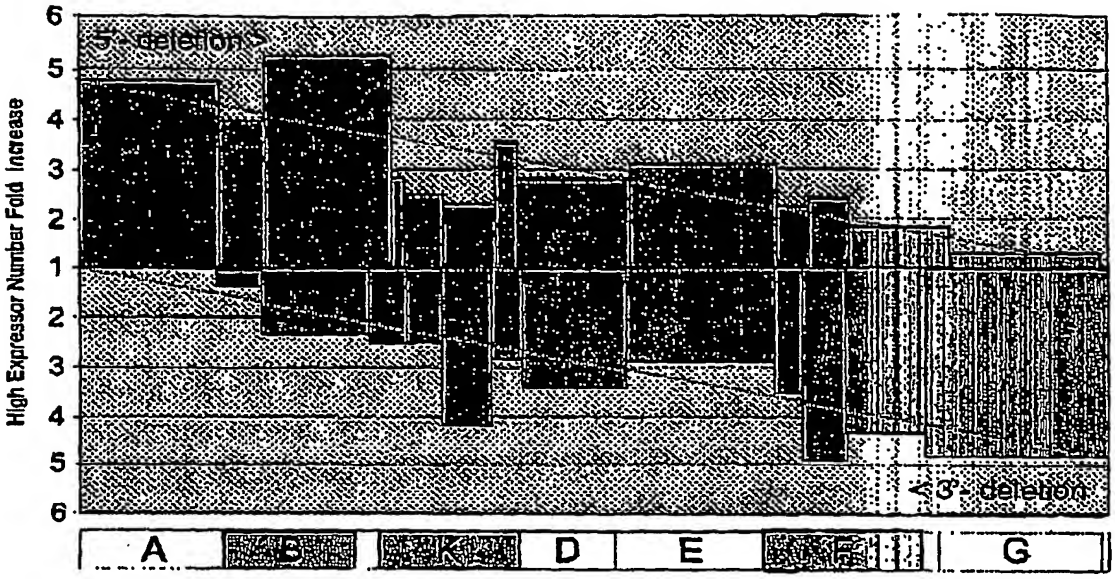

5'- deletion
3'- deletion
High Expressor Number Fold Increase
6
5
4
3
2
1
2
3
4
5
6
A
B
K
D
E
F
G

FIG.7

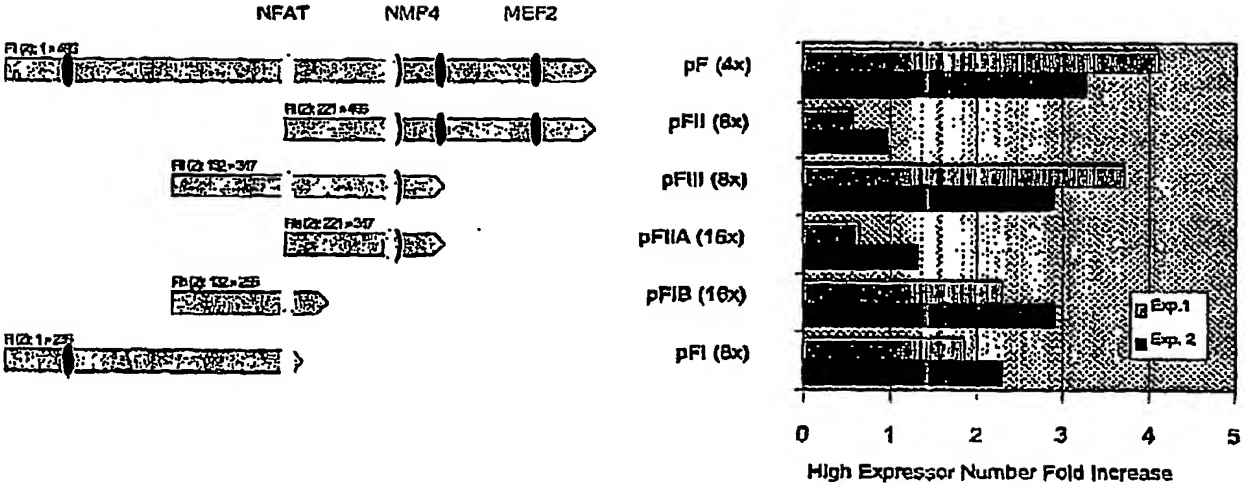

NFAT
NMP4
MEF2
pF (4x)
pFII (8x)
pFIII (8x)
pFIIA (16x)
pFIB (16x)
pFI (8x)
Exp.1
Exp.2
0
1
2
3
4
5
High Expressor Number Fold Increase

FIG.8

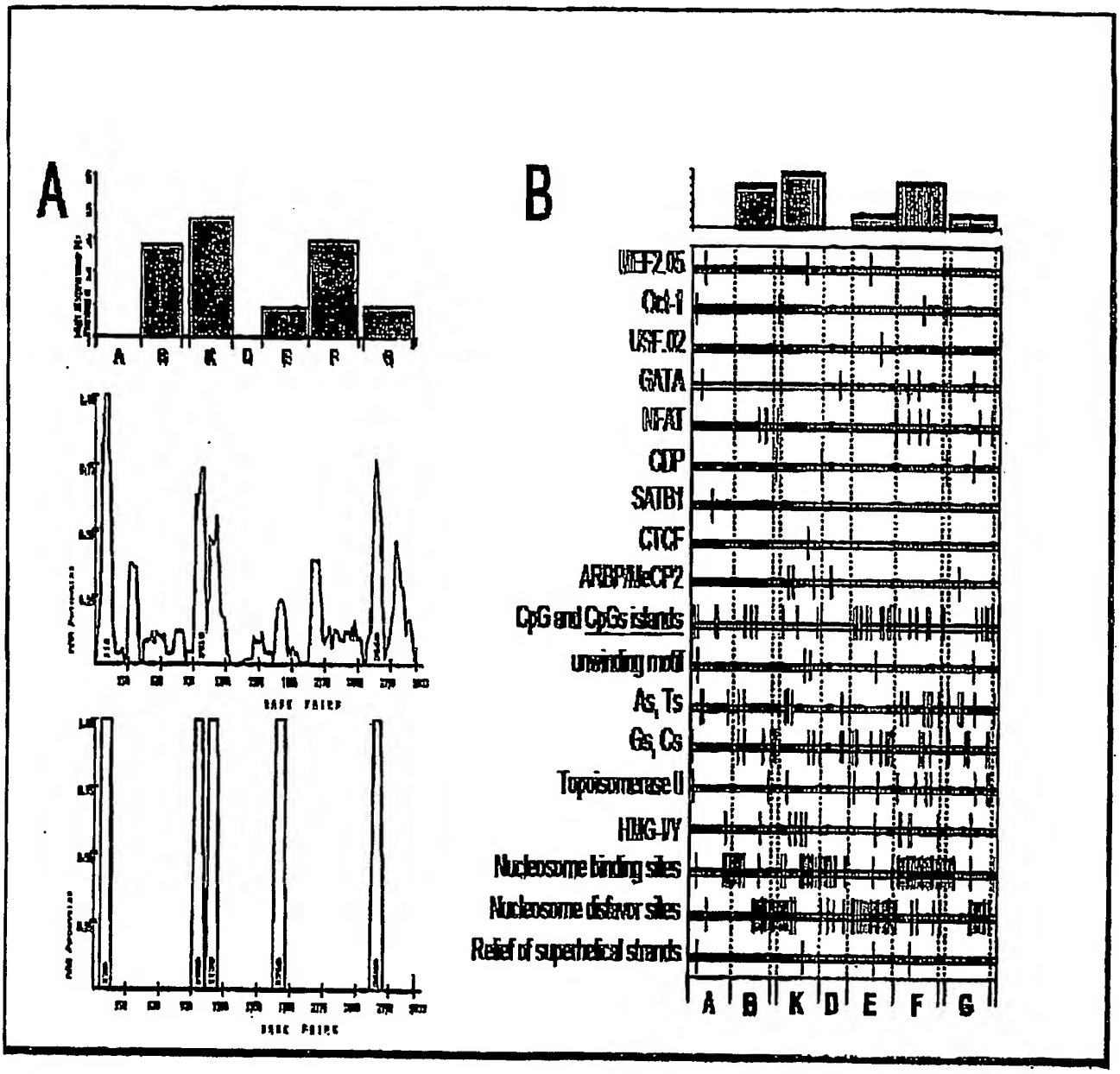
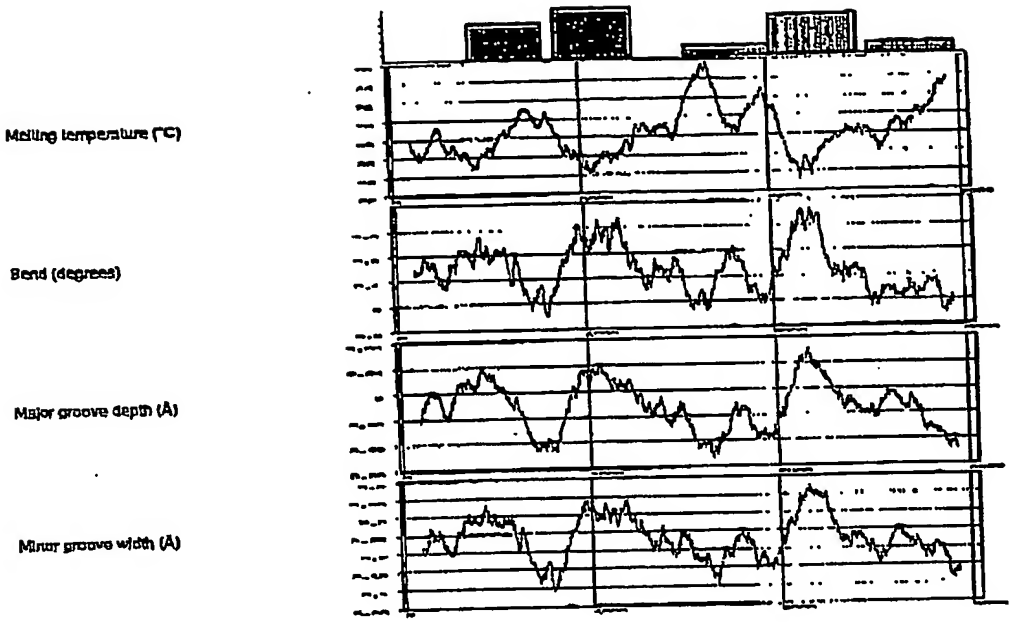

A
B
MEF2.05
Oct-1
USF.02
GATA
NFAT
CDP
SATB1
CTCF
ARBP/MeCP2
CpG and CpGs islands
unwinding motif
As, Ts
Gs, Cs
Topoisomerase II
HMG-IY
Nucleosome binding sites
Nucleosome disfavor sites
Relief of superhelical strands
A B K D E F G

FIG.9

(A)


Melting temperature (°C)
Bend (degrees)
Major groove depth (Å)
Minor groove width (Å)

(B)

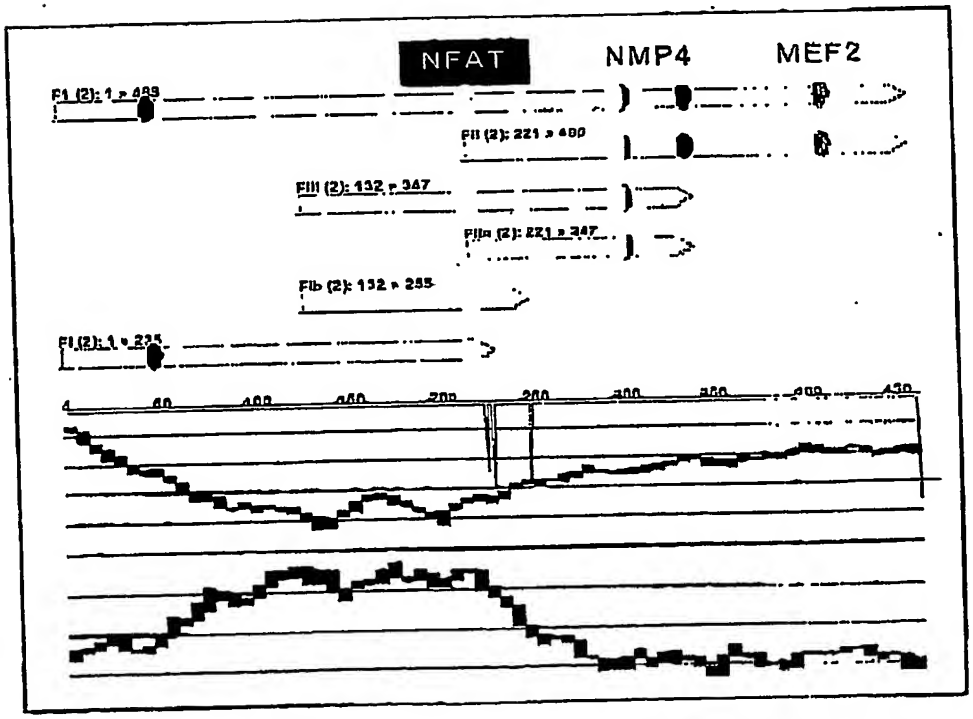

NFAT
NMP4
MEF2
FI (2): 1 » 489
FII (2): 221 » 489
FIII (2): 132 » 347
FIIIa (2): 221 » 347
FIb (2): 132 » 255
FI (2): 1 » 235

FIG.10

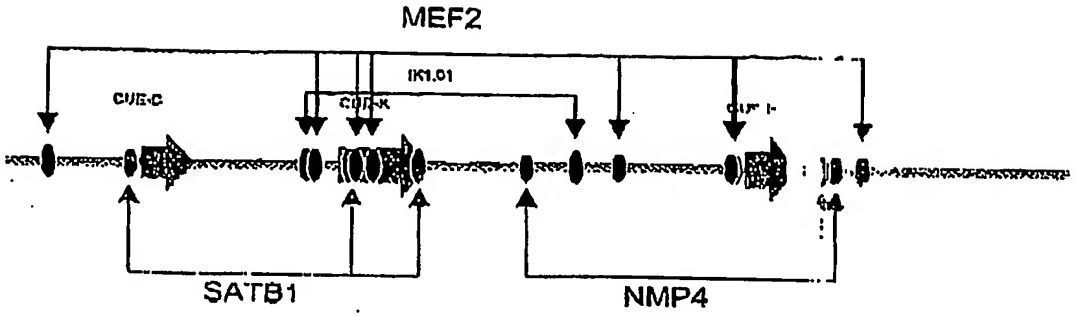

MEF2
IK1.01
CUE-C
SATB1
NMP4

FIG.11

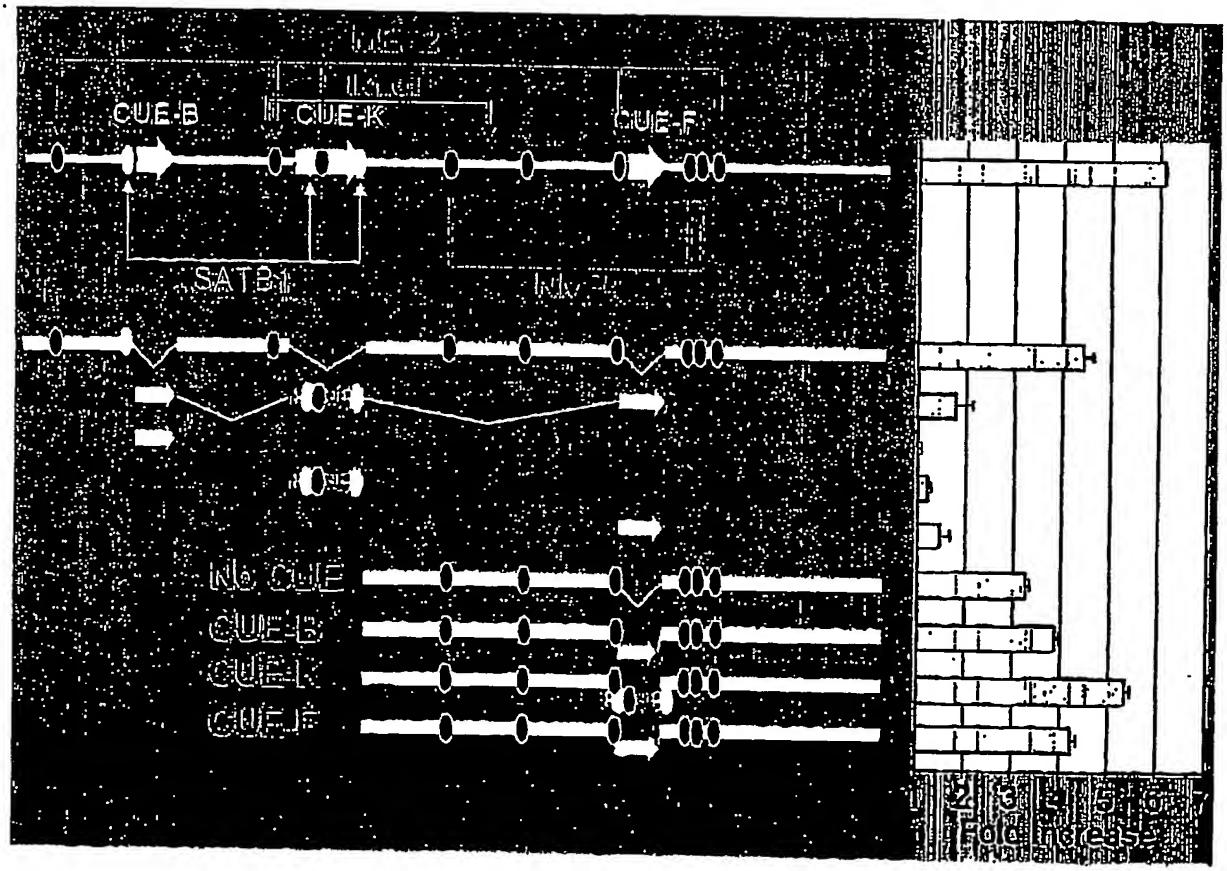

CUE-B
CUE-K
CUE-F
SATB1
No CUE
CUE-B
CUE-K
CUE-F
Fold Increase

FIG.12

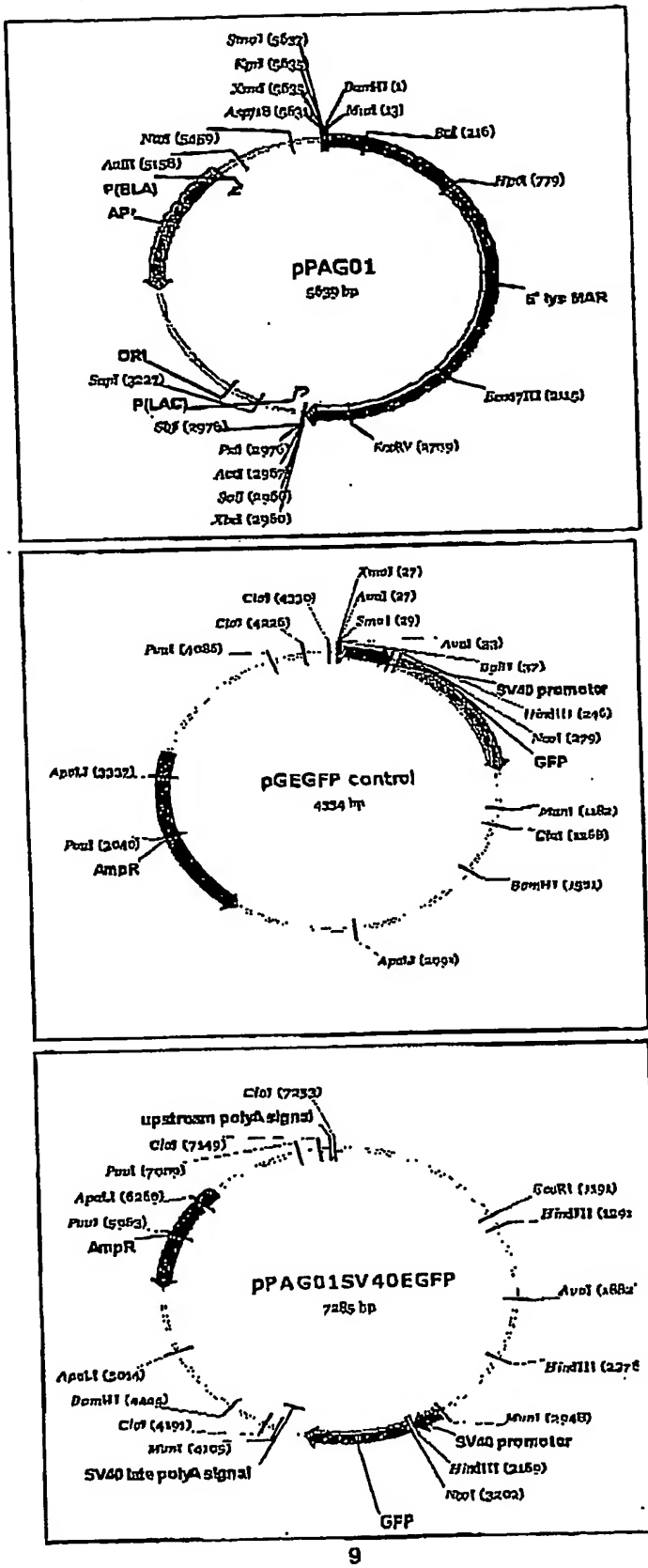

pPAG01
5639 bp
5' lys MAR
HpaI (779)
BclI (216)
NruI (5459)
AniII (5158)
P(BLA)
AP
ORI
SapI (3222)
P(LAC)
SbfI (2976)
PstI (2970)
AccI (2957)
SalI (2950)
XbaI (2950)
EcoRV (3709)
pGEGFP control
4334 bp
SV40 promoter
HindIII (246)
NcoI (279)
GFP
AmpR
BamHI (1931)
ApaLI (2094)
pPAG01SV40EGFP
7285 bp
upstream polyA signal
SV40 late polyA signal
HindIII (2159)
NcoI (3202)
AvaI (1682)
EcoRI (1594)

FIG.13

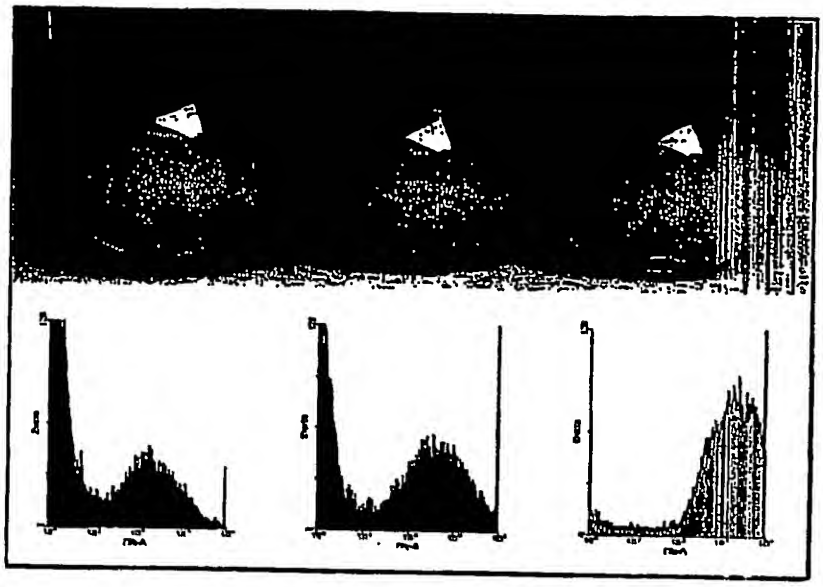
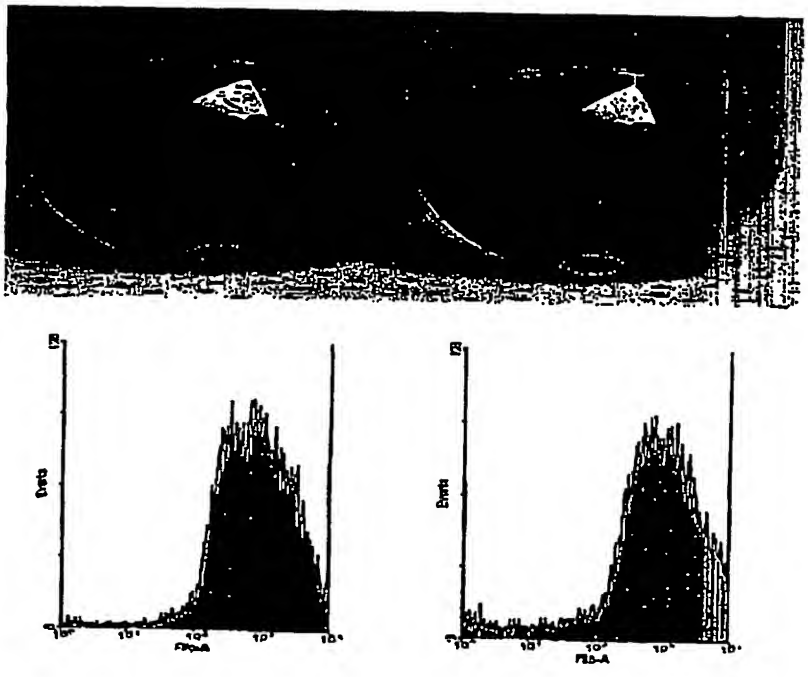


FIG.14



SEQUENCE LISTING

<110> Selexis S.A.

<120> High Efficiency Gene Transfer and expression inMammalian Cells by a Multiple Transfection Procedure of MAR Sequences

<130> SEL EP 003

<160> 23

<170> PatentIn version 3.1

<210> 1
<211> 320
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(320)
<223> MAR of human chromosome 1, nt from 36686 to 37008

<400> 1
ttatattatg ttgttatata tattatatta tgttattaga ttatattatg ttgttata
tt 60

atataataat attatattat atattatata ttatattata taatatataa taatatta
ta 120

taattatata ttacattata taatatataa taatattata taattatata ttacatta
ta 180

taatatataa taatattata taataatata taattatata atatataata atattata
ta 240

atattatata atattatata atatataaat atataataat atatattata ttatataa
ta 300

gtatataata ttatataata
320

<210> 2
<211> 709
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(709)

<223> MAR of human chromosome 1, nt from 142276 to 142984

<400> 2
tacaatatat tttctattat atatattttg tattatatat aatatacaat atattttc
ta 60

ttatatataa tatattttgt attatatata ttacaatata ttttgtatta tataatat
at 120

aatacaatat ataatatatt gtattatata ttatataata caatatatta tatattgt
at 180

tatatattat atataatact atataatata ttgtattata tattatatat aatactat
at 240

aatatatttt attatatatt atatataata caatatataa tatattgtat tataatac
aa 300

tgtattataa tgtattatat tgtattatat attatatata atacaatata taataata
ta 360

ttataatata taataataat ataatataat aataatatat attgtattat atattata
ta 420

atacaatata taatatattg tattatatat attttattac atataatata taatacat
ta 480

tataatatat tttgtattat atataatata ttttattatg tattatagat aatatatt
tt 540

attatatatt atatataata caatatataa tatattttgt attgtatata atatataa
ta 600

caatatataa tatattgtat tatatataat attaatatat tttgtattat atatttat
at 660

tttatattat aattatgttt tgcattatat atttcatatt atataiacc
709

<210> 3
<211> 409
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(409)
<223> MAR of human chromosome 1, nt from 1368659 to 1369067

<400> 3
tacacataaa tacatatgca tatatattat gtatatatac ataaatacat atgcatat
ac 60

attatgtata tatacataaa tacatatgca tatacattat gtatatatac ataaatac
at 120

atgcatatac attatgtata tatacataaa tacatatgca tatacattat gtatatat
ac 180

ataaatacat atgcatatac attatgtata tatacataaa tacatatgca tatacatt
at 240

gtatatatac ataaatacat atgcatatat tatatacata aattatatta tatacata
at 300

acatatacat atattatgtg tatatataca taaatacata tacatatatt atgtgtat
at 360

atacatgata catatacata tattatgtat atatatacat aaatacata
409

<210> 4
<211> 394
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(394)
<223> MAR of human chromosome 1, nt from 2839089 to 2839482

<400> 4
tatgtatata tacacacata tgtatatata cacacatatg tatatacgta tatatgta
ta 60

tatacacaca tatgtatata cgtatatatg tatatataca cacatatgta tatacgta
ta 120

tatgtatata tacacacata tgtatatacg tatatatgta tatatacaca catatgta
ta 180

tatgtatata tacacacata tgtatatacg tatatatgta tatatacaca catgtgta
ta 240

tatatataca catatgtata tatgtatata tacacacata tgtatatatg tgtatgta
ta 300

tatacacaca tatgtatata tacacatata tatgtatata tacacacata cttatata
ta 360

cacatatata tgtatatata cacatatgta taca
394

<210> 5
<211> 832
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(832)
<223> MAR of human chromosome 1, nt from 1452269 to 1453100

<400> 5
tatattacta tatatacaat atacatatta ctatatatac catgtattac tatatata
tc 60

tactatatat attactatat atacaaaata tatattacta tatatacaat atacatat
ta 120

ctatatatac catatattac tatatatatc tactatatat attactatat atacaaaa
ta 180

tatattacta tatatactat atattactgt atatacaata tatattacta tatatata
ct 240

atatattact atatatacac tatatattac tatatataca caatatatat attactat
at 300

atacacaatg tatataacta tatatacaat atatattact atatatacta tatatatt
ac 360

tatacatact atatattact ctatatatac aatatatata ttacaatata tactacat
at 420

tactacatat actttatata ttactatata tactatatat tactgtatat acaatata
ta 480

ttactaaata tacacaatat atattactat atatacacaa tatatatatt actatata
ta 540

cacattatat atgactatat atacacacta tatatattac tatatataca caatatat
aa 600

ctatatatac acagtataca tattactata tatacacaat atatatatta ctatatat
ac 660

actatatatt actatatata cacaatatat attactctat gtatacacta tatatatt
ac 720

tatatataca gaatatatat aactatatat acactatatt actatatata ctatatat
ta 780

ctatatgtac tatatatatt actatatata ctatatatta ctatatatac ac
832

<210> 6
<211> 350
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(350)
<223> MAR of human chromosome 1, nt from 831495 to 831844

<400> 6
aatatataat atataaatat taatatgtat tatataatat atattaatat attatatt
at 60

attactatat aaataatatt aatatattat attaaaatat taataaatat atcatatt
aa 120

atattatatt aattaaatat taataaatat attatattaa tatatttata tattaaac
ct 180

ataacatatg catatactta tttatatata acatgcatgt acttatttat atatacaa
ta 240

tatatttata tattatataa tatattatat gtatttatat attatatatc atatatta
ta 300

tgtatttata tattatatat catataatat atatatttat attatatata
350

<210> 7
<211> 386
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(386)
<223> MAR of human chromosome 1, nt from 1447225 to 1447610

<400> 7
acatttaatt taattatata ctgctatata taattaaatc tatatatcta tataactt

at 60

aatttatttt aatttaatta tatatactat atagttatat atacatatat gtaattat
at 120

atagtataat tatagtatat atgtatatat aatgtaagta aatatatagt atatattt
at 180

atatactata tatttataca tatgtcttta tatatactaa tatatataca catatgta
at 240

atgtacatat ggcatatatt ttatagtgta tatatacata tatgtaatat atatagta
at 300

atgtaaatat atagtacata tttaattata tggtaatata tacacatata tgtaatat
gt 360

gtattatagt acatatttta tagtat
386

<210> 8
<211> 585
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(585)
<223> MAR of human chromosome 1, nt from 4955365 to 4955949

<400> 8
atacacacat atacacatat gtacgtatat atactatata tacacacata tacacata
tg 60

tacgtatata tactatatat acacacatat acacatatgt acgtatatat actatata
ta 120

cacacatata cacatatgta cgtatatata ctatatatac acacatatac acatatgt
ac 180

gtatatatac tatatataca cacatataca catatgtacg tatatattat atatacac
ac 240

atatacacat atgtacgtat atatactata tatacacaca tatacacata tgtacgta
ta 300

tatactatat atacacacat atacacatat gtacgtatat atactatata tacacaca
ta 360

tacacatatg tacgtatata tactatatat acacacatat acacatatgt acgtatat

at 420

actatatata cacacatata cacatatgta cgtatatata ctatatatac acacatat
ac 480

acatatgtac gtatatatac tatatataca cacatataca catatgtacg tatatata
ct 540

atatataccc atacacatac gtatatacgt acatatatat acgta
585

<210> 9
<211> 772
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(772)
<223> MAR of human chromosome 1, nt from 5971862 to 5972633

<400> 9
agtaaacata tatatagtaa atatatatag tgtatatata gtaaatatat atagtgca
ta 60

tatatagtgc atatatatag tgtatatata gtaaatatat agtgtatata tatagtaa
at 120

atatatagtg tatatatagt aaatatatat agtaaatata tatatactat atatagta
aa 180

tatatatata ctatatatag taaatatata tatagtatat atatagtaaa tatatata
ta 240

gtatatatat agtaaatata tatatagtat atatatagta aatatatata tagtatat
at 300

agtaaatata tatagtatat atatagtaaa tatatatata gtatatatat agtaaata
ta 360

tatatagtat atatatagta aatatatata tagtatatat atagtaaata tatatagt
at 420

atatatagta aatatatata gtatatatat agtaaatata tatagtatat atatagta
aa 480

tatatataca ctgtatatat atagtaaata tatatacact gtatatatat agtaaata
ta 540

tatacactgt atatatatag taaatatata tacactgtat atatatagta aatatata

ta 600

cactgtatat acatagtaaa tatatataca ctgtatatac atagtaaata tatataca
ct 660

gtatatacat agtaaatata tatacactgt atatacatag taaatatata tacagtgt
at 720

atacatagta aatatatata cagtgtatat acatagtaaa tatatataca gt
772

<210> 10
<211> 304
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(304)
<223> MAR of human chromosome 1, nt from 6221897 to 6222200

<400> 10
atatataata tatataatta tattatatat aatatataat atatataatt atattata
ta 60

ttatatataa tatattatat attatatata taatatatat tatatattaa atatatat
ta 120

tatatataat atatattata tattaaatat atattatata tataatatat attatata
ta 180

atatatataa tatatattat atatatatta tatattatat atatatatta tatatata
ta 240

atatatataa tatatattat atataatata tattatatat atataatata tataatat
at 300

atta
304

<210> 11
<211> 311
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(311)
<223> MAR of human chromosome 1, nt from 9418531 to 9418841

<400> 11
tatatataat atttatatat aatattcatg tatttatata taaatattta tatattta
ta 60

tataaatatt tatatattta tatataaata tttatatatt tatatataat atttatac
at 120

tatatataat atttatatat tatatataat atttatatat aatatttata tattatat
at 180

aatatttata tatttatatg tataatatat attttatata tgtatgtata atatatat
tt 240

tatatatgta tgtataatat attttatata tgtatgtata atatattatt atatataa
ta 300

tataatttat a
311

<210> 12
<211> 302
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(302)
<223> MAR of human chromosome 1, nt from 15088789 to 15089090

<400> 12
atataatata tatattatat atataaatat atataaatat ataacatata tattatat
at 60

aaatatatat aaatatataa catatatatt atatatataa atatatataa atatataa
ca 120

tatatattat atatataaat atatataaat atataacata tatattatat atataaat
at 180

atattatata tttatatata taatatatat aaatatataa tatatattta tatatata
at 240

atatataaat atataatata tatatttata tataatatat ataaatatat aatatata
at 300

at
302

<210> 13
<211> 461
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(461)
<223> MAR of human chromosome 1, nt from 6791827 to 6792287

<400> 13
tatataatat atattatata tacacatata taatatatat tatatataca catatata
at 60

atatattata tatacacata tataatatat attatatata cacatatata atatatat
ta 120

tatatacaca tatataatat atattatata tacacatata taatatatat tatatata
ca 180

catatataat atatattata tatacacata tataatatat attatatata cacatata
ta 240

atatatatta tatatacaca tatgtaatat atattataca cacacatata atatatat
ta 300

tatacacata tataatatat attatatata catatataat atatattata tatacaca
ta 360

tataatatat attatatata cacatatata atatatatta tatatacaca tataatat
at 420

aatatataca catatataat atatatatta tatatgcaca t
461

<210> 14
<211> 572
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(572)
<223> MAR of human chromosome 1, nt from 163530 to 164101

<400> 14
atattataat tatatatatt atatataatt atataaaata tatattataa ttatatat
at 60

tttatataat atatatatta taattaatat attatatata atatatatat tatatata
at 120

atatatatta tatatattat atataatata tataatatat ataatatata atataata
ta 180

tatattatat ataatatata atatatataa tatattataa tataatatat ataatata
ta 240

atataatata tataatatat aatataatat ataatatata atatatataa tatataat
at 300

aatatataat atatataata tataatataa tatataatat atataatata ttataata
ta 360

atatatataa tatataatat aatatatata atatataata taatatataa tatataat
at 420

atatttaata tatttattaa ttatttgtta tatatttatt aatatataat atataata
ta 480

tttaatatat tataactata tattatatta taattatata tattatatat atacaatt
at 540

aattatatat tatatatact tataatatat at
572

<210> 15
<211> 357
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(357)
<223> MAR of human chromosome 1, nt from 1842332 to 1842688

<400> 15
tatatctata tatatctata tatatataat atagataata tctatatata taatatag
at 60

aatattatct atatataata tagataatat tatctatata taatatagat aatattat
ct 120

atatataaaa ttatattata tctatatata ttatatatat aaaattatat tatatcta
ta 180

tataatatag ataatatcta tatataaata gataatatct atatatataa tatagata
tt 240

atctatatta tagatataga taatattatc tatattatag atattatcta tatataat
at 300

agataatatt atctatatta tatatataat atatctatat tatctataat attatct
357

<210> 16
<211> 399
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(399)
<223> MAR of human chromosome 1, nt from 2309560 to 2309958

<400> 16
attatatata atatatatta tatattatat atatcaagca gcagatataa tatataat
at 60

atataatata tataatatat attgtatatt atataatata taatatatat aatatata
tt 120

gtatattata taatatataa tatatataat atatattgta tattatataa tatataat
at 180

atataatata tattgtatat tatataatat ataatatatg taatatatta tgtaatat
at 240

tatataatat atattatata ttatatataa tatatattat atatuatata tattacat
aa 300

tatattacat atattacgta atatatgtta tatattacat ataatatata acatatat
ta 360

cgtaatatat gtaatatatt acatataata tatacatta
399

<210> 17
<211> 394
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(394)
<223> MAR of human chromosome 1, nt from 2231759 to 2232152

<400> 17
atatatactt ataaattata tacttatata tacttataaa ttatatactt atatatac
tt 60

ataaattata tacttatata tacttataaa ttatatactt atatatactt ataaatta
ta 120

tacttatata tacttataaa ttatatactt atatatactt ataaattata tacttata
ta 180

tacttataaa ttatatactt atatataatt ataaattata tacttatata taattata
aa 240

ttatatactt atatataatt ataaattata tacttatata taattataaa ttatatac
tt 300

atatataatt ataaattata tacatatata taattataaa ttatatacat atataatt
at 360

aaattatata catatataat tataaattat atac
394

<210> 18
<211> 387
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(387)
<223> MAR of human chromosome 1, nt from 7406524 to 7406910

<400> 18
tatattatat ataatatata ttatatataa tataaataat atatattata tataatat
at 60

aaataatata taatatataa ataatatata atatataata tataaataat atataata
ta 120

taacatataa ataatatata taatatataa ataatatata taatatataa ataatata
ta 180

taatatataa aaatatataa tatataatac atatataaat aatatattat attatata
tg 240

atacataata tattatatat aatatattat atgatacata atatattata tagaatat
at 300

tatatgatac ataatatatt atatagaata tattatatga tacataatat attatatg

at 360

acataatata ttatatataa tatatta
387

<210> 19
<211> 370
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(370)
<223> MAR of human chromosome 1, nt from 9399572 to 9399941

<400> 19
catatataca tatatacaca tatatacaca tatatataca catacatatg tacacata
ta 60

tatacacata tgtatacaca tatatacaca tatatacaca catatataca catatata
ca 120

cacatatata cacatatata cacatatata cacatataca catatataca catatata
ca 180

tatatacaca tatatataat atacacacat atatatacac atatatacac acatatat
ac 240

acatatatac acatatatat acacatatat acacatatat acatatatac acatatat
at 300

acatatatac acatatatac atatatacac atatatacat atatacacac atatatac
ac 360

atacatatac
370

<210> 20
<211> 377
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(377)
<223> MAR of human chromosome 1, nt from 12417411 to 12417787

<400> 20

attatatata atacatataa ttatatattt atatataaat tataataaat acatataa
tt 60

atatatttat atataaatta tatataataa atacatataa ttacatatat ttataaat
ta 120

taataaatac atataattac atatatttat atatgaatta tatataataa atacatat
aa 180

ttatatatat ttatatgtag attatatata aatatatata atttatatat ataataat
at 240

atataattta tatatataat tatatatata ataaatatat ataatttata tatataat
ta 300

tatatataat aaatatataa taatatatat aatttatata tataattata tatataat
aa 360

atatatataa tttatat
377

<210> 21
<211> 1524
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)...(1524)
<223> MAR of human chromosome 1, nt from 1643307 to 1644830

<400> 21
tataaatata tataaatata taaatatata taaatatata aatatatata aatatata
ta 60

aatatataaa aatatataaa tatatataaa tatatataaa tatataaaaa cataaaaa
ta 120

tatataaata tatataaata tataaaaata tataaatata taaatatata aaaatata
ca 180

aatatataaa tatatacata aatatatata aatatatata aatatataaa aatatata
ta 240

aatatataaa tatatataaa tatatataaa tatatataaa tatataaaaa tatatata
aa 300

tatataaata tataaaaata tatataaata tataaatata taaatatata taaatata
ta 360

aatatataaa taaatataag tatttatgaa tatatatgaa tatataaata tataaaaa
at 420

atatataaat atataaatat atataaatat ataaatatat acatatatac atatataa
at 480

aaataaatat aagtatttat gaatatatat gaatatataa atatataaaa aatatata
ta 540

aatatataaa tatatataaa tataaatata taaaaatata taaaaatata tataaata
ta 600

taaatatata taaatatata aatatatata aatatatata aatatataaa tatatata
aa 660

tatatataaa tatataaata tataaatata tataaatata tataaatata taaatata
ta 720

aatataaata tataaatata tataaatata tataaatata taaatatata taaatata
ta 780

taaatatata taaatatata taaatatata aatatatata aatatatata taaatata
ta 840

taaatatata aatatataaa tatataaaaa tatataacaa tatataaata tatataaa
aa 900

tatataacaa tatataaata taaatatata taaaaatata taacaatata taaatata
aa 960

tatatataaa tatataaata taaatataaa aaatatatat aaatatataa atatatat
aa 1020

atatataaat gtataaatat atataaaaat atataacaat atataaatat ataaatat
at 1080

aacaatatat aaatatataa aaatatataa caatatataa atataaatat atataaaa
at 1140

atataacaat atataaatat aaatatatat ataaatatat aaatataaat ataaaaaa
ta 1200

tatataaata tataaatata tatataaata tatataaata tataaatgta taaatata
ta 1260

taaatatata aatatataaa aatatataaa tatatataaa tatatataaa tatataaa
ta 1320

taaatatata aatatatata aatatataaa tataaatata taaacatata taaatata
ta 1380

taaataaaca tatataaaga tatataaaga tataaagata tataaatata taaatata

ta 1440

aagatatata aatatataaa gatatataaa tatataaaga tatataaata tataaaga
ta 1500

tataaatata atatataaat atat
1524

<210> 22
<211> 664
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(664)
<223> MAR of human chromosome 1, nt from 1398763 to 1399426

<400> 22
acacatatat atataaaata tatatatata cacacatata tataaaatat atatatat
ac 60

acacatatat ataaaatata tatatacaca catatatata aaatatatat atacacac
at 120

atatataaaa tatatatata cacacatata tataaaatat atatatacac acatatat
at 180

aaaatatata tatacacaca tatatataaa atatatatat acacacatat atataaaa
ta 240

tatatataca cacatatata taaaatatat atatacacac atatatataa aatatata
ta 300

tacacacata tatataaaat atatatatac acacatatat aaaatatata tatacaca
ca 360

tatataaaat atatatatac acatatatat aaaatatata tatacacata tatataaa
at 420

atatatacac acatatatat aaaatatata tatacacaca tatatataaa atatatat
at 480

acacatatat ataaaatata tatatacaca tatatataaa atatatatat atacacat
at 540

atataaaata tatatacaca catatatata aagtatatat atacacacat atatataa
aa 600

tatatatata cacatatata taaaatatat atatacacat atatataaaa tatatata

ta 660

caca
664

<210> 23
<211> 1428
<212> DNA
<213> Homo sapiens

<220>
<221> misc_binding
<222> (1)..(1428)
<223> MAR of human chromosome 2, nt from 17840365 to 17841792

<400> 23
aatttattat atattatata ttatatatat tatatatatt atatattata tatattat
at 60

atattatata ttatatatat tatatattat atatttatat ataatatata tctaatat
at 120

atattagata taatatatat ctaatatata tatattttat atataiaata tatctcta
at 180

atatatattt tatatgtata taatatatct ctaatatata tatattttat atgtatat
aa 240

tatatctcta atatatatat tttttatata taatatatct ctaalatata tattttat
at 300

atataatata tatctaatat atataatata tatattagat atatataaaa tatatatg
at 360

atatttatta tatatataat atataatata taatatatat attatattat atacatat
at 420

attatataca atatatatta tatatatttt atatacatta tatattatal atatttta
ta 480

tacaatatat attatatatt ttatatacaa tatatattat atatattttа tattttta
ta 540

tacaatatat attatatata ttttatatat aatatatatt atatatattt tatataat
at 600

atattatata tattttatat ataatatata ttatataaat tatatataat atatatta
ta 660

ataaattata atatttttta tatatataat atgtatttta tatataatat attataat

at 720

atattttata tataatatat tataatatat attttatata taatatatta taatatat
at 780

tttatattat aatatattat aatatatatt ttatatataa tatattataa tatatatt
tt 840

atatataata tattataata tatattttat atataatata ttataatata tatattat
aa 900

tatatatttt atatataata tattatcata tatatattaa atatatattt tatatata
at 960

atattataat atatatatta taatatatat tttatatata atatattata atatatat
at 1020

tataatatat attttatata taatatatta taatatatat tttatatata atatatta
ta 1080

atatatattt tatatataat atattataat atatatttta tatataatat aatatata
tt 1140

ttatatataa tatattataa tatatatttt atatataata tattataata tatatttt
at 1200

atataatata ttataatata tattttatat ataatatatt ataatatata ttttatat
at 1260

aatatattat aatatatatt ttatatataa tatattataa tatatatttt atatataa
ta 1320

tattataata tatattttat atataatata ttaattaaat ttattaattt attaatta
tt 1380

aatatttatt atattattaa ttaataatat ataaattatt aatatata
1428